REPRESENTATIONS + LIE ALGEBRA = EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Neural Networks (GNNs) are powerful yet remain 'black box' models, and their use in high-stakes domains requires explanation methods that are both mathematically principled and semantically faithful. Existing factual explainers identify influential subgraphs via discrete combinatorial search, while counterfactual methods perturb the latent embedding without geometric constraints, frequently producing out-of-distribution (OOD) artifacts that are infeasible or structurally invalid. We address both limitations by framing GNN reasoning as symmetry transformation on a learned latent manifold. To this end, we introduce GLIDE: Graph LIe Decomposition Explainer. We model the GNN latent space as a smooth manifold acted upon by a continuous Lie group G, and we learn a basis $\{A_k\}$ for the associated Lie algebra $\mathfrak{g}$. Factual explanations correspond to the label-stabilizer orbit (group actions generated by the subalgebra $\mathfrak{h} \subset \mathfrak{g}$ that leave the predicted class invariant). Counterfactual explanations correspond to transverse flows generated by the complement $\mathfrak{m}$, which traverse the decision boundary along the shortest algebraic path. We prove that this decomposition is reductive, guarantee that all generated trajectories remain within a bounded neighborhood of the data manifold, establish that orthogonal generator bases produce locally independent counterfactual directions, and show that winner-take-all sparsity recovers the unique minimal-norm counterfactual generator. Empirical evaluation on synthetic (BA-2Motifs) and biochemical (MUTAG, PROTEINS, NCI1, AIDS) benchmarks demonstrates that our framework achieves high-quality factual and counterfactual explanations with strict manifold proximity, outperforming all baseline methods.

1 INTRODUCTION

Graph Neural Networks (GNNs) have demonstrated strong performance on non-Euclidean, relational data, achieving state-of-the-art results in drug discovery dejun et al. (2020); Luong & Singh (2024); Xiong et al. (2020), molecular simulation Reiser et al. (2022); Sanchez-Gonzalez et al. (2020), social network analysis Amelkin et al. (2017); Bogdanov et al. (2013), and neuroscience Bullmore & Sporns (2009); Cui et al. (2023); Lin et al. (2022). However, the "black box" nature of GNNs limits deployment in high-stakes domains where interpretability is needed for trustworthiness and domain validation of the reasoning underlying the GNN's prediction.

The explainable AI (XAI) literature contains two complementary, but currently disjoint, paradigms. Factual explainers identify a minimal subgraph that is most influential for the model's prediction Luo et al. (2020); Miao et al. (2022); Seo et al. (2024); Ying et al. (2019); Yuan et al. (2021). Counterfactual explainers complement this by seeking a minimally perturbed graph for which the model's prediction changes Chen et al. (2023); Kosan et al. (2022); Lucic et al. (2022); Ma et al. (2022). The main challenge here is proximity. Existing methods frequently produce invalid out-of-distribution (OOD) artifacts; interpretable counterfactuals must reside strictly on the data manifold.

A challenge shared by both factual and counterfactual paradigms is how to enforce and interpret geometric structure on the latent space $\mathcal{Z}$ that is simultaneously (i) compatible with the classifier's decision boundary, (ii) rich enough to express both label-preserving and label-altering transformations, and (iii) closed under composition such that multi-step reasoning remains on the manifold. Three concrete gaps in existing methods illustrate why current approaches fall short. First, subgraph selectors (Ying et al. (2019); Luo et al. (2020); Miao et al. (2022); Seo et al. (2024)) return one representative structure per instance, with no native object for the full, continuously-varying set of structures that would receive the same label. Second, counterfactual generators (Ma et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2022); Chen et al. (2023); Kosan et al. (2022)) enforce proximity only as a soft penalty inside a VAE/diffusion/random-walk objective, with no closed-form bound relating the transformation's parameters to the resulting drift off the manifold. Third, margin-based reasoning more broadly gives a scalar distance to the boundary but no notion of a minimal, non-redundant direction.

Lie groups satisfy (i)–(iii) by construction. A Lie group G acting smoothly on $\mathcal{Z}$ gives a group of differentiable, invertible transformations whose infinitesimal generators (elements of the associated Lie algebra $\mathfrak{g}$) provide a linear, global basis system for reasoning about motion in $\mathcal{Z}$. The Lie algebra admits a canonical decomposition in which the generators that leave the classifier's decision invariant form a subalgebra $\mathfrak{h}_z \subset \mathfrak{g}$ (the label-stabilizer algebra), while the complement $\mathfrak{m}_z = \mathfrak{g} \ominus \mathfrak{h}_z$ generates flows that cross it. This decomposition is precisely what makes a minimal, non-redundant crossing direction well-posed (Theorem 1), where scalar margins alone cannot. Since the transformation is parameterized via the matrix exponential, composing these flows admits an explicit bound relating generator norm and coefficient magnitude directly to drift off the manifold (Theorem 3, Eq. 8). These guarantees require a differentiable group structure on $\mathcal{Z}$, illustrating the need for the Lie-theoretic framework to provide solutions to each of these classical limitations.

We propose GLIDE. We unify factual subgraph identification and geometrically valid counterfactual traversal under a Lie-algebraic framework by modeling the GNN latent space as a manifold with a continuous Lie group action G. The factual orbit of z under class y corresponds to the stabilizer subgroup $H_z \leq G$, whose generators form the factual subalgebra $\mathfrak{h}_z \subset \mathfrak{g}$ and induce class-preserving deformations. Counterfactuals arise from the orthogonal complement $\mathfrak{m}_z = \mathfrak{g} \ominus \mathfrak{h}_z$, whose flows cross the decision boundary transversely. Generated via the exponential map $\exp : \mathfrak{g} \to G$, these trajectories are differentiable, invertible, and intrinsically manifold-constrained, eliminating OOD drift in standard counterfactual methods.

We establish a formal theoretical foundation for our approach. We prove the decomposition is reductive (Theorem 1), derive a manifold proximity bound for Lie trajectories as a regularizable constraint (Theorem 3, Theorem 5), show orthogonal generators yield locally independent counterfactual directions (Proposition 2), characterize the adjoint regularization ensuring subalgebra consistency (Proposition 3), show winner-take-all sparsity recovers the minimal-norm counterfactual generator (Theorem 4), and prove factual learning precedes counterfactual learning (Theorem 6). Our primary contributions are as follows:

(1) **GLIDE, a symmetry-theoretic GNN explainer.** We model factual and counterfactual reasoning as continuous symmetry transformations on a learned Lie group manifold, achieving state-of-the-art performance among baselines.

(2) **Factual/counterfactual disentanglement.** We prove that the factual and counterfactual subalgebras form a reductive pair under adjoint regularization, yielding a clean separation of invariant and variant reasoning directions.

(3) **Manifold proximity guarantees.** We derive explicit bounds ensuring that Lie-generated counterfactuals remain in a close neighborhood of the data manifold, solving OOD drift.

(4) **Sparsity and minimality.** We show that temperature-scaled winner-take-all sparsity over orthogonal generators recovers the unique minimal-norm counterfactual transformation, grounding conciseness in an algebraic optimality principle.

2 Related Work

Factual GNN explainability. Post-hoc methods such as GNNExplainer Ying et al. (2019) and PG-Explainer Luo et al. (2020) train separate masks on a frozen GNN to maximize mutual information between a subgraph and the model prediction. Ante-hoc methods such as GSAT Miao et al. (2022) and PGIB Seo et al. (2024) incorporate information-theoretic bottlenecks directly into GNN training. None of these methods also supply counterfactual reasoning; GLIDE subsumes factual explanation as the stabilizer orbit and supplies the counterfactual complement in a single unified model.

Counterfactual GNN explainability. GCFExplainer Kosan et al. (2022) and COMRECGC Fournier & Medya (2025) use importance-weighted random walks to generate global counterfactuals. CLEAR Ma et al. (2022) decouples counterfactual generation into a VAE-based latent traversal; and D4Explainer Chen et al. (2023) uses diffusion. These methods do not guarantee that

the traversed latent path corresponds to a valid graph. Our Lie flow approach constructs trajectories that are natively confined to the learned group manifold, eliminating OOD explanations.

Symmetry and equivariance in graph learning. Equivariant GNNs such as SE(3)-Transformers Fuchs et al. (2020) and E(n)-GNNs Satorras et al. (2022) hard-code physical symmetries to achievce invariance by design, but do not identify which symmetry drives the decision. We recover symmetry generators from the latent space during training rather than enforcing them as inductive bias.

Symmetry discovery via Lie algebra parameterization. Augerino Benton et al. (2020) learns augmentation policies as linear combinations of Lie algebra bases, while LieGAN Yang et al. (2023) and SymmetryGAN Desai et al. (2022) use adversarial training to discover label-preserving symmetries. LieGG Moskalev et al. (2023) extracts Lie bases from discriminator polarization matrices, and LaLiGAN Yang et al. (2024) extends this to nonlinear symmetries via latent linearization. (Dehmamy et al., 2021) enforce discovered symmetries via weight sharing, and (Romero & Lohit, 2023) learn task-relevant partial equivariances. A novelty of our work is supervised disentanglement. Whereas existing methods recover a single symmetry, we partition the algebra into label-preserving and label-altering components, a theoretically grounded (Theorem 1) and necessary decomposition for explainability. For a detailed comparison with these methods please see Appendix L.

3 THEORY

We develop a Lie-algebraic framework where factual and counterfactual explanations are geometric objects. Background on Lie groups/ algebras is in Appendix D; Appendix C lists all notation.

3.1 LIE-THEORETIC UNDERSTANDING OF EXPLANATION

Definition 1 (Label-Stabilizer Subgroup)**.** *Let G be a group acting smoothly on $\mathcal{Z}$ and let $z \in \mathcal{Z}$ have predicted label $\hat{y}(z)$. The label-stabilizer of z is $H_z = \{g \in G : \hat{y}(g \cdot z) = \hat{y}(z)\}$. The connected label-stabilizer H_z^0 is the connected component of H_z containing the identity e.*

The connected component H_z^0 forms a submanifold of G near the identity e, whose Lie algebra $\mathfrak{h}_z = T_e H_z^0$ is well-defined and closed under the Lie bracket, provided the logit $y(z) = [h_\phi(z)]_c$ has nonzero gradient at z (i.e., the decision boundary is smooth near z; see Proposition 1, Assumption 1). We enforce this structure via adjoint regularization (Proposition 3); see Section 5.

3.1.1 FACTUAL EXPLANATION.

A factual explanation is a minimal subgraph $\mathcal{G}_{Sub} \subseteq \mathcal{G}$ (or its latent image z_{Sub}) that is maximally informative of the prediction $\hat{y}(z)$. We redefine factual explanations in the Lie-theoretic context.

Definition 2 (Factual Explanation)**.** *The* factual explanation *of z under $h_\phi \circ f_\theta$ is the stabilizer orbit*

$$\mathcal{O}_{\hat{y}}(z) = \{\exp(X_{\text{fact}}) \cdot z \mid X_{\text{fact}} \in \mathfrak{h}_z\}, \tag{1}$$

i.e., the set of embeddings reached by label-preserving Lie flows. The saliency $s_k = \left\|\nabla_X t_{f,k}\right\|_2$ is the explanatory subgraph, induced by the dominant generator A_{f,k^*}, where $k^* = \arg\max_k s_k$.

This definition generalizes classical mutual-information approaches, which identify a single explanatory subgraph $\mathcal{G}_{\text{Sub}}^*$ Miao et al. (2022); Seo et al. (2024); Yuan et al. (2021). Instead, Definition 2 characterizes the full equivalence class of subgraphs assigned to class $\hat{y}(z)$, of which $\mathcal{G}_{\text{Sub}}^*$ is one representative. The orbit $\mathcal{O}_{\hat{y}}(z)$ thus provides a complete characterization of factual reasoning.

Proposition 1 (Lie Algebra of the Label-Stabilizer)**.** *Let $c = \hat{y}(z)$ be the predicted class at z, and let $y(z) = [h_\phi(z)]_c$ denote the logit for class c, which is C^1 by smoothness of h_ϕ. Let a Lie group G act smoothly on $\mathbb{R}^D$ with Lie algebra $\mathfrak{g}$, and define the label differential $\ell_z : \mathfrak{g} \to \mathbb{R}$ by*

$$\ell_z(A) := \nabla_z y(z)^\top (Az), \quad A \in \mathfrak{g}.$$

Suppose the following three conditions hold at z: (1) Non-degeneracy: $\nabla_z y(z) \neq 0$; (2) Faithfulness: The evaluation map $\rho_z : \mathfrak{g} \to \mathbb{R}^D$, $A \mapsto Az$, is injective. (3) Local transitivity: $\rho_z(\mathfrak{h}_z) = T_z\mathcal{Z}_y$, where $T_z\mathcal{Z}_y = \ker\big(\nabla_z y(z)^\top\big)$.

Then: (i) *$\mathfrak{h}_z$ is a Lie subalgebra of $\mathfrak{g}$;* (ii) $\mathfrak{h}_z = \ker \ell_z$*;* (iii) $\dim \mathfrak{h}_z = \dim \mathfrak{g} - 1$.

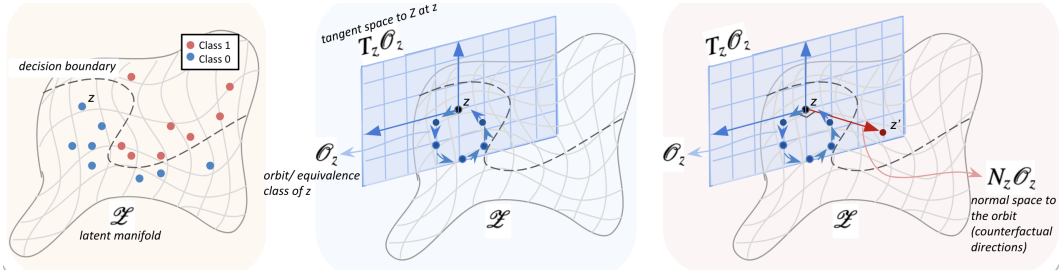

Figure 1: GLIDE latent space. Left Panel (Latent Manifold): latent space $\mathcal{Z}$ with graph embeddings, of Class 0 (blue) and Class 1 (red) separated by a dashed decision boundary. Middle Panel (Tangent Space & Factual Orbits): the tangent space $T_z\mathcal{O}_z$ at a given latent point z, with label-preserving factual flows along the orbit $\mathcal{O}_z$ that stay within the equivalence class. Right Panel (Counterfactual Direction): the normal space $N_z\mathcal{O}_z$, showing trajectory to z' across the decision boundary.

Proposition 1 shows that classifier infinitesimal symmetries coincide with the decision manifold's tangent space ($\mathfrak{h}_z = \ker \ell_z$ by local transitivity). Generators $A \in \mathfrak{h}_z$ induce label-preserving motions (Az tangent to level sets, $\nabla_z y(z)^\top Az = 0$), decomposing the Lie algebra into a tangent space with a 1D normal direction. This canonical transverse direction in $\mathfrak{g}/\mathfrak{h}_z$ generates single-element minimal counterfactuals (Theorem 4; see Appdx. H.2 and N.1 for proofs and empirical validations).

3.1.2 Counterfactual Explanation.

A counterfactual explanation seeks $z' \in \mathcal{Z}$ with $\hat{y}(z') \neq \hat{y}(z)$ and minimal edits from z. The key difficulty is enforcing $z' \in \mathcal{Z}$, as optimization in $\mathbb{R}^D$ can leave the data manifold and produce OOD embeddings. We re-frame counterfactual explanations in Lie-theoretic terms.

Definition 3 (Counterfactual Explanation)**.** *The counterfactual explanation of z is the set of embeddings reachable via label-changing Lie flows achieving y_{cf}, namely, the transverse orbit:*

$$\mathcal{T}_{y_{\text{cf}}}(z) = \left\{ \exp(X_{\text{cf}}) \cdot z \mid X_{\text{cf}} \in \mathfrak{m}_z,\ \hat{y}(\exp(X_{\text{cf}}) \cdot z) = y_{\text{cf}} \right\}, \tag{2}$$

The explanation is the minimal-displacement element

$$z^* = \exp(t_{k^*} A_{cf,k^*}) \cdot z, \text{with } k^* = \arg\min_k \| \exp(t_k A_{cf,k}) \cdot z - z \|_2, \tag{3}$$

with input-space saliency $s_k = \left\| \nabla_X\, t_{cf,k} \right\|_2$ over X, induced by A_{cf,k^*}.

The transverse orbit $\mathcal{T}_{y_{\text{cf}}}(z)$ is a submanifold of the homogeneous space G/H_z^0 (Corollary 1), with tangent space $\mathfrak{m}_z$ at H_z^0, formalizing that counterfactual reasoning moves z across the decision boundary. Since every point in $\mathcal{T}_{y_{\text{cf}}}(z)$ is reached via the exponential map from $\mathfrak{m}_z$, the trajectory from $z \to z^*$ is smooth, invertible, and bounded within the neighborhood guaranteed by Theorem 3.

3.2 Factual and Counterfactual Decomposition

Definitions 2 and 3 establish geometric notions of reasoning. Proposition 1 identifies the Lie algebra of label-preserving directions with the kernel of the label differential, $\mathfrak{h}_z = \ker \ell_z$. Under non-degeneracy, $\nabla_z y(z) \neq 0$, this kernel is a codimension-one subspace of $\mathfrak{g}$ (Lemma 1), corresponding to the tangent space of the decision level set at z. This induces the decomposition:

$$\mathfrak{g} = \mathfrak{h}_z \oplus \mathfrak{m}_z, \tag{4}$$

where $\mathfrak{m}_z := \mathfrak{h}_z^\perp$ by the Frobenius inner product. The subspace $\mathfrak{h}_z$ captures label-preserving directions, while $\mathfrak{m}_z$ is the unique first-order direction that moves z across the decision boundary.

Theorem 1 (Reductive Decomposition and Factored Group Action)**.** *If the decomposition equation 4 satisfies the reductive conditions*

$$[\mathfrak{h}_z, \mathfrak{h}_z] \subseteq \mathfrak{h}_z, \qquad [\mathfrak{h}_z, \mathfrak{m}_z] \subseteq \mathfrak{m}_z, \tag{5}$$

then $(\mathfrak{g}, \mathfrak{h}_z)$ is a reductive pair, and locally near $e \in G$ every $g \in G$ admits the unique factorization

$$g = \exp(X_{\text{fact}}) \cdot \exp(X_{\text{cf}}), \qquad X_{\text{fact}} \in \mathfrak{h}_z,\ X_{\text{cf}} \in \mathfrak{m}_z. \tag{6}$$

We give the proof in Appendix H.1. The factual ($\mathfrak{h}_z$) and counterfactual ($\mathfrak{m}_z$) subalgebras depend solely on the decision boundary geometry of f_θ and the data distribution. Canonical bases for $\mathfrak{m}_z$ (non-unique up to $O(\dim \mathfrak{m}_z)$) are fixed via SVD of the learned generators. Theorem 1 is exact when $L_{adj} = 0$ (Equation 39) and gains accuracy as training drives $L_{adj} \to 0$.

4 Extended Theoretical Guarantees

Theorem 2 (Factual Orbits Foliate the Class Manifold). *Let $\mathcal{Z}_y = \{z \in \mathcal{Z} : \hat{y}(z) = y\}$ denote the class manifold for label y. The collection of factual orbits defines a smooth foliation of $\mathcal{Z}_y$:*

$$\mathcal{F} = \big\{\mathcal{O}_y(z)\big\}_{z \in \mathcal{Z}_y}, \quad \mathcal{O}_y(z) = \{\exp(X_{\text{fact}}) \cdot z : X_{\text{fact}} \in \mathfrak{h}_z\}, \tag{7}$$

*Furthermore, (i) **Completeness.** For any $z' \in \mathcal{Z}_y$ sufficiently close to z, there exists $X_{\text{fact}} \in \mathfrak{h}_z$ such that $z' = \exp(X_{\text{fact}}) \cdot z$. The factual explanation is therefore complete and it captures all local label-preserving representations of z, not merely a finite subgraph representative point; and (ii) **Decision Boundary as Leaf Space.** The leaves of the transverse foliation $\mathcal{T}_{y_{\text{cf}}}(z)|_{z \in \mathcal{Z}_y}$ coincide with the decision boundary $\partial \mathcal{Z}_y$, i.e., it is the union of counterfactual orbit leaves.*

We provide the proof in Appendix H.4. Part (i) establishes the benefit of the orbit-based factual explanation over classical subgraph selection. A subgraph $\mathcal{G}^*_{Sub}$ is a single representative of the leaf $\mathcal{O}_y(z)$, whereas Definition 2 characterizes the entire leaf. Part (ii) gives a novel geometric definition of the decision boundary as the leaf space of the counterfactual foliation. Empirical validation is given in Section 7, with visualization of the decision boundary as the union of counterfactual orbit leaves in Figure 7. This foliation structure is a derived consequence (via Frobenius theorem from bracket-closure in Prop. 1), not an imposed assumption. A simpler formalization would not necessarily produce a foliation, confirming the Lie-subalgebra as the correct underlying model.

Corollary 1 (Homogeneous Space of Counterfactual Directions). *Under the conditions of Theorem 1, the manifold of counterfactual transformations is the reductive homogeneous space*

$$G/H^0_z = \{gH^0_z \mid g \in G\},$$

whose tangent space at the identity coset H^0_z is naturally identified with $\mathfrak{m}_z \cong \mathfrak{g}/\mathfrak{h}_z$.

Equation 6 cleanly factors reasoning modes where $\exp(X_{\text{fact}})$ moves z within its label-stabilizer orbit, identifying invariant symmetries within the model, and $\exp(X_{\text{cf}})$ moves z across the decision boundary along the geodesic of the homogeneous space G/H^0_z to yield minimal perturbations.

4.1 Manifold Proximity Guarantee

Lie group structure explicitly bounds OOD drift, a key issue in counterfactual explanation Chen et al. (2024); Kosan et al. (2022); Ma et al. (2022). Theorem 3 gives a hard constraint on OOD drift.

Theorem 3 (Manifold Proximity Bound). *Let $\{A_k\}_{k=1}^K$ be the counterfactual generator basis with $\|A_k\|_F \le M \ \forall k$, and let $t \in \mathbb{R}^K$ be the coefficient vector. Define $z' = \exp\Big(\sum_{k=1}^K t_k A_k\Big)\, z$. Then:*

$$\|z' - z\|_2 \le \|z\|_2 \left(e^{M\|t\|_1} - 1\right). \tag{8}$$

For tolerance $\delta > 0$, z' satisfies $\|z' - z\|_2 \le \delta$ whenever $\|t\|_1 \le \frac{1}{M} \log\Big(1 + \frac{\delta}{\|z\|_2}\Big)$.

We provide the proof in Appendix H.5. The bound depends only on the generator norms M and the coefficient magnitude $\|t\|_1$, both of which are regularized during training. The low-rank factorization $A_k = U_k V_k^\top$ additionally constrains $\|A_k\|_F \le \|U_k\|_F \|V_k\|_F$, providing an architecture-level upper bound on M. Theorem 5 (Appendix H.7) strengthens Theorem 3 to a dynamic guarantee. The proximity bound tightens monotonically over training as generators converge to their fixed point A^*_k and the coefficient network g_ψ learns increasingly efficient boundary-crossing directions, driving $M^*\|t\|_1 \to 0$ while maintaining $M^* > 0$ and high CF Fidelity. This provides a formal justification for monitoring the proximity loss as a convergence diagnostic. Results in Table 2 confirm the lowest CF Proximity with GLIDE. Empirical validation is in Section 7 and Figure 8.

5 METHODS

GLIDE decomposes the tangent space at latent state $z \in \mathcal{Z}$ as $T_z\mathcal{Z} = T_z\mathcal{O}_z \oplus N_z\mathcal{O}_z$, where $T_z\mathcal{O}_z$ is the orbit tangent space (factual directions spanned by $\{A_{f,k}z\}$) and $N_z\mathcal{O}_z$ is the normal bundle (counterfactual directions spanned by $\{A_{cf,k}z\}$). Factual transformations move z along the decision manifold, leaving predictions unchanged, and counterfactual transformations move z across the decision boundary, changing the class label, providing a complete geometric view of reasoning. We depict the latent space decomposition in Figure 1, and our pipeline diagram in Figure 2.

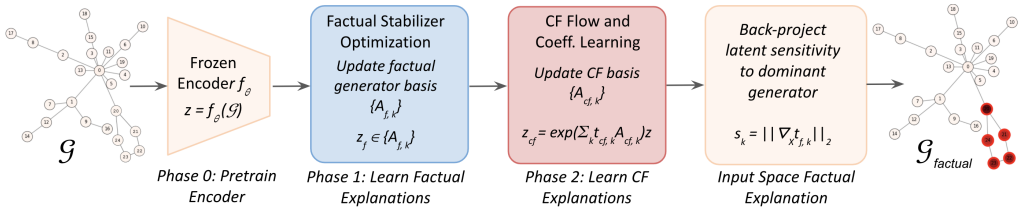

Figure 2: Phase 0 (Pretrain f_θ): The input graph $\mathcal{G}$ is encoded via frozen GNN f_θ to yield the latent representation $z = f_\theta(\mathcal{G})$. Phase 1 (Factual Explanations): Optimizes the factual generator basis $\{A_{\text{f},k}\}$ to model label-stabilizing transformations ($z_f \in \{A_{\text{f},k}\}$). Phase 2 (CF Explanations): Computes counterfactual flows and learns coefficient via exponential mapping ($z_{\text{cf}} = \exp(\sum_k t_{\text{cf},k} A_{\text{cf},k})z$) to update the counterfactual basis $\{A_{\text{cf},k}\}$. Input Space Factual Explanation: Back-projects latent sensitivities to the dominant generator, yielding subgraph saliency map ($\mathcal{G}_{\text{factual}}$).

5.1 SYMMETRY-THEORETIC LATENT SPACE LOW-RANK PARAMETERIZATION

Let $f_\theta : \mathcal{G} \to \mathcal{Z} \subset \mathbb{R}^D$ be a GNN encoder and $h_\phi : \mathcal{Z} \to \mathbb{R}^C$ a classifier. G is a matrix Lie group acting on $\mathcal{Z}$. We parameterize its Lie algebra $\mathfrak{g}$ with K basis generators $\{A_k\}_{k=1}^K$, $A_k \in \mathbb{R}^{D\times D}$. To ensure bounded transformations, each generator is factored as $A_k = U_k V_k^\top$ with $U_k, V_k \in \mathbb{R}^{D\times r}$ and rank $r \ll D$. This constrains $\|A_k\|_F \le \|U_k\|_F \|V_k\|_F$, providing architecture-level control over the proximity bound in Theorem 3. A group action on z is $z' = \exp\left(\sum_{k=1}^K t_k A_k\right) z$, where $t \in \mathbb{R}^K$ is a coefficient vector predicted by a network g_ψ.

5.2 GLIDE ALGORITHM

We include the pseudocode algorithm for GLIDE in Appendix E, and discuss the factual and counterfactual learning in Sections 5.2 and 5.2, respectively. The encoder/classifier is trained in Stage 0, then frozen for Phases 1–2 where only the generator matrices and coefficient network are updated.

Phase 1: Factual Orbit Learning We learn K factual generators $A_{f,k}k = 1^K$ whose flows remain in the label-stabilizer orbit of z by minimizing $\mathcal{L}_{\text{ce}}(\text{softmax}(h_\phi(z_f)), y)$, where $z_f = \exp(\sum_k \alpha_k A_{f,k})z$ with $\alpha_k \sim \mathcal{N}(0, I)$. We enforce tangent alignment and Lie closure (Proposition 1), along with orthogonality and proximity (losses defined in Appendix I). The full factual objective is: $\mathcal{L}_{total}^{(f)} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{tan}\mathcal{L}_{tan} + \lambda_{closure}\mathcal{L}_{closure} + \lambda_{prox}\mathcal{L}_{prox} + \lambda_{ortho}\mathcal{L}_{ortho}$.

Phase 2: Counterfactual Reasoning via Lie Flow With $\{A_{f,k}\}$ fixed at convergence of Phase 1, we train K counterfactual generators $\{A_{cf,k}\}$ together with a coefficient network $g_\psi(z, y_{cf}) \to t_{cf} \in \mathbb{R}^K$ that predicts the transformation magnitudes required to flip the prediction to the target class y_{cf}. Propositions 2 and 3 motivate $\mathcal{L}_{\text{ortho}}$ and $\mathcal{L}_{\text{adj}}$, respectively. The sparsity and margin-based label flip losses are detailed in Appendix J to yield the final counterfactual objective, $\mathcal{L}_{\text{total}}^{(cf)} = \lambda_{\text{margin}}\mathcal{L}_{\text{margin}} + \lambda_{\text{prox}}\mathcal{L}_{\text{prox}} + \lambda_{\text{entropy}}\mathcal{L}_{\text{entropy}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{adj}}\mathcal{L}_{\text{adj}}$. All loss terms are ablated in Appendix K and shown to be necessary.

5.3 SPECTRAL THEORY OF TWO-PHASE TRAINING

Algorithm 1's two-phase protocol fixes the stabilizer subgroup before exploring the transverse complement. Theorem 6 (Appendix H.8) establishes this spectral rationale via Hessian curvature, which accelerates factual over counterfactual subspace convergence. The spectral gap

$\Delta = \lambda_{\min}(H|_{\mathfrak{h}_z}) - \lambda_{\min}(H|_{\mathfrak{m}_z})$ governs phase separation and training interference. Large Δ yields brief, distinct phases; small Δ leads to overlap where joint training is less harmful (Section 7).

5.4 Minimality and Sparsity of Counterfactual Generators

A key requirement for interpretable explanations is concision. The counterfactual should be expressible by a single, identifiable reasoning direction rather than a combination of all generators. Theorem 4 below (proof in Appendix H.6) formalizes the sparsity constraint recovering the algebraically optimal counterfactual. With Proposition 2, the generator k^* is guaranteed to be the semantically unique, minimum direction of change in model reasoning. See Section 7 for empirical validation.

Theorem 4 (Optimal Generator Selection under One-Hot Restriction)**.** *Under the assumptions of Proposition 1,* $\dim \mathfrak{m}_z = 1$*. Let* $\{A_{cf,k}\}_{k=1}^K$ *be Frobenius-orthogonal counterfactual generators with projections* $a_k := \langle A_{cf,k}, m\rangle_F$ *onto the unit normal* m *of* $\mathfrak{m}_z$*. Among one-hot coefficient vectors* $t = t_{k^*}e_{k^*}$ *satisfying the first-order label-flip constraint* $|t_{k^*}a_{k^*}| \geq \theta$*, the minimum-latent-displacement solution is attained at:* $k^* = \arg\max_k \frac{|a_k|}{\|A_{cf,k}z\|_2}$*, with* $|t^*_{k^*}| = \frac{\theta}{|a_{k^*}|}$.

6 Results

We evaluate factual and counterfactual baseline methods on synthetic and real-world benchmarks. GLIDE achieves state-of-the-art performance on all metrics, datasets, and explanation modes.

6.1 Factual Explanation Performance

We compare GLIDE to factual explainers GSAT Miao et al. (2022) and PGIB Seo et al. (2024), and symmetry-discovery methods LieGAN Yang et al. (2023) and Augerino$^+$ Benton et al. (2020). Both symmetry-discovery baselines are adapted for explainability (Appendix L.1) to operate in the latent space of our frozen encoder f_θ and are evaluated with the same metrics as our method, ensuring that any difference in explanation quality reflects the method performance rather than representational capacity. Results for PGIB and GSAT are reported in Table 1. Results on LieGAN and Augerino$^+$ are given in Appendix L.3; they require extended discussion since they are symmetry-discovery methods, not inherently explainers. As shown in Table 1, GLIDE outperforms all baselines. Appendix G visualizes the exact identification of ground-truth explanations on BA-2Motifs via GLIDE.

Table 1: Factual explanation and classification performance. Results are mean $\pm$ std over 10 runs. GLIDEoutperforms all baselines for classification and explanation.

Method	Metric	BA_2Motifs	MUTAG	PROTEINS	NCI1	AIDS
GLIDE	Acc (↑)	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.97 \pm 0.07}$	$\mathbf{0.72 \pm 0.02}$	$\mathbf{0.80 \pm 0.05}$	$\mathbf{0.93 \pm 0.01}$
	Fid+ (↓)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.04 \pm 0.01}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.05 \pm 0.00}$
	Fid− (↑)	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.98 \pm 0.10}$	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.78 \pm 0.04}$	$\mathbf{0.95 \pm 0.01}$
GSAT	Acc (↑)	0.99 ± 0.01	0.91 ± 0.03	0.68 ± 0.04	0.69 ± 0.02	0.83 ± 0.03
	Fid+ (↓)	0.47 ± 0.11	0.34 ± 0.02	0.71 ± 0.21	0.54 ± 0.6	0.49 ± 0.01
	Fid− (↑)	0.61 ± 0.09	0.42 ± 0.32	0.73 ± 0.05	0.62 ± 0.05	0.59 ± 0.07
PGIB	Acc (↑)	0.63 ± 0.17	0.90 ± 0.01	0.72 ± 0.02	0.66 ± 0.02	0.82 ± 0.01
	Fid+ (↓)	0.49 ± 0.15	0.59 ± 0.20	0.60 ± 0.13	0.52 ± 0.16	0.28 ± 0.04
	Fid− (↑)	0.83 ± 0.16	0.75 ± 0.08	0.64 ± 0.02	0.45 ± 0.12	0.51 ± 0.02

Collapse of symmetry-discovery baselines. Table 10 in Appendix L.3 reveals that both LieGAN and Augerino$^+$ learn near-degenerate generators. We have $\|A_{f,k^*}\|_F < 0.05$ for both methods, producing orbit radii $\|g(\boldsymbol{t})z - z\|_2 < 0.001$ and group elements satisfying $\|\exp(\sum_k t_k A_{f,k^*}) - I\|_F < 0.001$. The reported $\mathrm{Fid}^- \approx 1.0$ therefore reflects the trivial fact that the identity transformation preserves the class label, not that a meaningful factual explanation has been found. This collapse arises because neither method optimizes an objective requiring a non-trivial orbit. LieGAN's adversarial loss and Augerino$^+$'s invariance loss are both minimized by $A_{f,k^*} \to 0$. GLIDE avoids this by construction as the margin loss $\max(0, \gamma - \Delta_f(z', y))$ is zero only when z' is confidently inside the class region, which requires the orbit to move z by a non-trivial amount.

6.2 Counterfactual Explanation Performance

We compare against CLEAR Ma et al. (2022) (VAE latent traversal), D4 Chen et al. (2023) (diffusion), and GCFExplainer Kosan et al. (2022) and COMRECGC Fournier & Medya (2025) (random walks for global counterfactuals). We report CF Validity (rate at which z' attains the target label) and CF Proximity ($\|z_{cf} - z\|_2$, measuring OOD drift, computed only over valid counterfactuals to measure the latent space proximity for all baselines; no graph-space edit distance is used). All entries are directly comparable on a shared scale. Implementation details are in Appendix M. GLIDE outperforms all baseline methods for CF Validity and CF Proximity (Table 2), and the latent displacement $\mathcal{L}_{\text{prox}}$ remains tightly bounded, consistent with the guarantee of Theorem 3. Baselines have much worse CF Proximity, confirming the OOD drift problem identified in Section 3.1.

Table 2: Counterfactual explanation performance. Results are mean ± std over 10 runs. GLIDE surpasses all baselines in both CF Validity and Proximity metrics.

Method	Metric	BA_2Motifs	MUTAG	PROTEINS	NCI1	AIDS
GLIDE	CF Validity (↑)	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.91 \pm 0.14}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.97 \pm 0.05}$
	CF Prox (↓)	$\mathbf{2.74 \pm 1.94}$	$\mathbf{2.07 \pm 0.21}$	$\mathbf{1.30 \pm 1.84}$	$\mathbf{1.38 \pm 0.12}$	$\mathbf{3.11 \pm 0.31}$
CLEAR	CF Validity (↑)	0.80 ± 0.01	0.99 ± 0.00	0.62 ± 0.00	0.82 ± 0.16	0.91 ± 0.01
	CF Prox (↓)	7.95 ± 1.03	8.97 ± 1.03	14.61 ± 7.24	393.40 ± 343.98	7.93 ± 1.16
GCFExplainer	CF Validity (↑)	0.92 ± 0.01	0.93 ± 0.00	0.89 ± 0.01	0.88 ± 0.05	0.94 ± 0.01
	CF Prox (↓)	12.68 ± 6.33	37.60 ± 1.71	47.84 ± 3.20	38.69 ± 0.82	16.97 ± 2.06
COMRECGC	CF Validity (↑)	0.52 ± 0.00	0.93 ± 0.00	OOM	0.87 ± 0.02	0.92 ± 0.00
	CF Prox (↓)	4.65 ± 0.00	10.83 ± 0.06	OOM	11.67 ± 0.15	3.86 ± 0.02
D4 Explainer	CF Validity (↑)	0.49 ± 0.01	OOM	OOM	OOM	OOM
	CF Prox (↓)	7.83 ± 2.54	OOM	OOM	OOM	OOM

We provide two additional ablations in Appendix M.7 showing that GLIDE outperforms CF generation via (1) raw gradient-ascent in latent space, and (2) an ungrounded low-rank perturbation without the exponential map or reductive-decomposition losses (Tables 11 and 12). This isolates the group-exponential parameterization and the reductive-decomposition/ closure losses, while holding the low-rank architecture fixed. Both perform substantially worse than GLIDE (details below), providing direct empirical support for the necessity of the Lie-theoretic structure.

6.3 Visualization of Factual Orbits and Counterfactual Trajectories

Figure 3 illustrates the geometric structure of the learned latent space. Factual flows remain within each class region, while counterfactual flows cross the decision boundary, confirming the orbit structure of Definitions 2 and 3. In the PCA visualization, Class 0 (blue) and Class 1 (orange) form distinct clusters, with counterfactual trajectories (red) crossing into the target class. The trajectories remain near the boundary, reflecting the proximity guarantee of Theorem 3 and the reductive decomposition of Theorem 1, which together yield the clean partition of latent transformations.

7 Empirical Validation of Theoretical Guarantees

In this section, we validate each theoretical guarantee in Section 3 with experimental evidence on BA-2Motifs. The theory itself makes no dataset-specific claims. It follows that the validation holds across datasets, each showing this behavior. For complete analysis across datasets, and full details on all validation protocol implementations, please see Appendix N.

Validation of Theorem 2: The Decision Boundary as a Counterfactual Leaf Space For each test embedding z we binary-search for the minimal $\alpha^* \in [0, 1]$ such that $\hat{y}(\exp(\alpha^* B) \cdot z) = y_{\text{cf}}$, where $B = \sum_k t_{cf,k} A_{cf,k}$. The boundary is recovered for all 200 test samples, all lying on $\partial \mathcal{Z}_y$, with $\bar{\alpha}^* = 0.198 \pm 0.049$. The recovered points form the decision boundary (Figure 7, left), confirming the foliation of Theorem 2 (see Appendix N.2 for full protocol).

Validation of Theorem 3: Manifold Proximity Guarantees The effective B-norm is 8.845 (mean z-norm 3.282). Since $\bar{\alpha}^* = 0.198$, the mean boundary displacement is $\approx 19.8\%$ of the full transformation (well within the bound of Theorem 3). No sample requires more than 51.4% (Figure

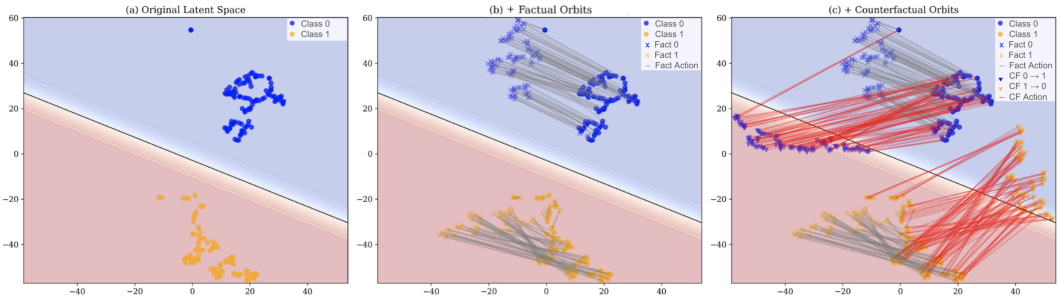

Figure 3: **Lie-theoretic decomposition of latent transformations via GLIDE.** (a) Learned latent space with decision boundary; circles denote classes 0 (blue) and 1 (orange). (b) $\times$ denotes factual explanations with gray transformations that lie in the stabilizer subalgebra $\mathfrak{h}_z$ ($Az \perp \nabla_z y(z)$), generating label-preserving orbits within each class. (c) Triangles denote counterfactuals with red trajectories that lie in the complementary subspace $\mathfrak{m}_z$, which is locally 1D and aligned with the boundary normal, yielding minimal boundary-crossing directions. This illustrates the reductive decomposition $\mathfrak{g} = \mathfrak{h}_z \oplus \mathfrak{m}_z$ and the sparsity result that a single counterfactual generator suffices.

7), confirming that counterfactual flows remain geometrically local latent space. We provide full analysis is in Appendix N.3, and Figure 8 shows decreasing proximity over epochs.

Validation of Theorem 4: Winner-Take-All Recovers the Minimum-Norm Generator The coefficient network concentrates all magnitude on a single generator A_0 with $\|A_0\|_F = 0.2447$ (mean coefficient 13.56; all other generators ≈ 0), and CF Fidelity $= 1.000$ is achieved using only this top-1 generator, verifying that a single Lie algebra direction suffices to characterize the full decision boundary crossing (Theorem 4). Appendix N.5 has the full protocol and discussion.

Validation of Theorem 6: Spectral Theory of Two-Phase Training We validate that two-phase training is spectrally necessary via the experiemnt detailed in Appendix N.4. Fid^- is maximized at the start of Phase 1 ($\tau_f \approx 0$), while CF Fidelity reaches 1.000 within $\tau_{cf} = 5$ epochs of Phase 2, confirming $\tau_f < \tau_{cf}$ (Figure 9). In the joint-training ablation (Figure 10), the same staircase ordering emerges without explicit phase separation, driven purely by the measured spectral gap ($\lambda_f = 1086 > \lambda_{cf} = 475$), directly validating the curvature inequality of Theorem 6.

8 Discussion, Limitations, and Future Directions

Discussion. We introduced GLIDE, a factual and counterfactual explainer grounded in Lie theory. A canonical reductive decomposition of a learned Lie algebra admits a single algebraic view where factual reasoning is motion within the label-stabilizer orbit, and counterfactual reasoning is flow across the decision boundary. Theorems 1–6 establish a formal geometric framework for GNN explanation with provable guarantees. GLIDE outperforms all baselines on five benchmarks and all metrics. Section 7 confirms the theory: decision boundary foliation, proximity control, spectral convergence ordering, and winner-take-all sparsity all emerge in the learned latent space.

Limitations. Three limitations merit acknowledgment. First, the approximation in Theorem 4 is first-order in $\|t\|_1$; for large transformations, higher-order corrections from the Baker-Campbell-Hausdorff (BCH) formula may shift the argmin. Second, while the low-rank factorization $A_k = U_k V_k^\top$ controls the proximity bound, the effect of rank r on expressiveness vs. manifold fidelity could be explored. Third, GLIDE gives factual explanations in the graph space, but, adding a decoder or generative pairing for discrete counterfactual edits in the graph space is open work.

Future directions. Natural extensions include: (i) replacing the ℓ_2 proximity loss with a Riemannian geodesic distance on G, improving the tightness of Theorem 3; (ii) generalizing from binary to multi-class counterfactuals by defining one complement subalgebra per target class (discussed in Appendix P); (iii) applying GLIDE to node-level explanation, where the group action must respect local neighborhood structure; (iv) investigating connections to topological data analysis, where the stabilizer orbits may be related to persistent homology classes of the decision boundary; (v) extending GLIDE to other domains, this method need not be restricted to the graph domain.

AI USE STATEMENT

We used generative AI tools minimally to help with basic tasks which do not involve intellectual contributions. We wrote all original text ourselves and occasionally used AI to edit prose for concision and clarity. We also used AI to help with the implementation of some supporting analysis and figure code. We did not use AI for any implementation of our main method. All experiments were designed and run by the authors, and their outputs were inspected by the authors, All numerical claims were checked against the saved outputs. The authors take full responsibility for the content, claims, code, figures, and artifacts of this work.

ETHICS STATEMENT

We believe that the ethical and societal implications of this work are consistent with those commonly encountered in machine learning research for scientific discovery, and we do not identify any specific risks that warrant additional discussion beyond standard considerations (i.e. responsible use).

REPRODUCIBILITY STATEMENT

All datasets used are available for download from the original source. All code needed to reproduce experiments and generate all figures within the paper is available in the supplementary material, and experimental setup information is outlined in the README. Follow the guidelines on the supplementary material README pages to reproduce the performance results, as well as the plots presented throughout the paper.

REFERENCES

Victor Amelkin, Francesco Bullo, and Ambuj K. Singh. Polar opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(11):5650–5665, 2017. doi: 10.1109/TAC.2017.2694341.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks, 2020. URL https://arxiv.org/abs/2010.11882.

Petko Bogdanov, Michael Busch, Jeff Moehlis, Ambuj K. Singh, and Boleslaw K. Szymanski. The social media genome: modeling individual topic-specific behavior in social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pp. 236–242. ACM, August 2013. doi: 10.1145/2492517.2492621. URL http://dx.doi.org/10.1145/2492517.2492621.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186—198, March 2009. ISSN 1471-003X. doi: 10.1038/nrn2575. URL https://doi.org/10.1038/nrn2575.

Jialin Chen, Shirley Wu, Abhijit Gupta, and Rex Ying. D4explainer: In-distribution gnn explanations via discrete denoising diffusion, 2023. URL https://arxiv.org/abs/2310.19321.

Zhuomin Chen, Jiaxing Zhang, Jingchao Ni, Xiaoting Li, Yuchen Bian, Md Mezbahul Islam, Ananda Mohan Mondal, Hua Wei, and Dongsheng Luo. Generating in-distribution proxy graphs for explaining graph neural networks, 2024. URL https://arxiv.org/abs/2402.02036.

Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: A benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2):493–506, February 2023. ISSN 1558-254X. doi: 10.1109/tmi.2022.3218745. URL http://dx.doi.org/10.1109/TMI.2022.3218745.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network, 2021. URL https://arxiv.org/abs/2109.07103.

Jiang dejun, Zhenxing Wu, Kim Hsieh, Chen Guangyong, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models, 09 2020.

Krish Desai, Benjamin Nachman, and Jesse Thaler. Symmetry discovery with deep learning. *Physical Review D*, 105(9), May 2022. ISSN 2470-0029. doi: 10.1103/physrevd.105.096031. URL http://dx.doi.org/10.1103/PhysRevD.105.096031.

Pavel Etingof. Lie groups and lie algebras, 2026. URL https://arxiv.org/abs/2201.09397.

Gregoire Fournier and Sourav Medya. Comrecgc: Global graph counterfactual explainer through common recourse, 2025. URL https://arxiv.org/abs/2505.07081.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020. URL https://arxiv.org/abs/2006.10503.

Mert Kosan, Zexi Huang, Sourav Medya, Sayan Ranu, and Ambuj Singh. Global counterfactual explainer for graph neural networks, 2022. URL https://arxiv.org/abs/2210.11695.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities, 2022. URL https://arxiv.org/abs/2210.01769.

Ana Lucic, Maartje ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks, 2022. URL https://arxiv.org/abs/2102.03322.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network, 2020. URL https://arxiv.org/abs/2011.04573.

Kha-Dinh Luong and Ambuj Singh. Application of transformers in cheminformatics. *Journal of chemical information and modeling*, 64, 05 2024. doi: 10.1021/acs.jcim.3c02070.

Jing Ma, Ruocheng Guo, Saumitra Mishra, Aidong Zhang, and Jundong Li. Clear: Generative counterfactual explanations on graphs, 2022. URL https://arxiv.org/abs/2210.08443.

Siqi Miao, Miaoyuan Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism, 2022. URL https://arxiv.org/abs/2201.12987.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators, 2023. URL https://arxiv.org/abs/2210.04345.

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry, 2022. URL https://arxiv.org/abs/2208.09481.

David W. Romero and Suhas Lohit. Learning partial equivariances from data, 2023. URL https://arxiv.org/abs/2110.10211.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks, 2020. URL https://arxiv.org/abs/2002.09405.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2022. URL https://arxiv.org/abs/2102.09844.

Sangwoo Seo, Sungwon Kim, and Chanyoung Park. Interpretable prototype-based graph information bottleneck, 2024. URL https://arxiv.org/abs/2310.19906.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xu tong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with graph attention mechanism. *Journal of medicinal chemistry*, 2020. URL https://api.semanticscholar.org/CorpusID:199572628.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery, 2023. URL https://arxiv.org/abs/2302.00236.

Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery, 2024. URL https://arxiv.org/abs/2310.00105.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks, 2019. URL https://arxiv.org/abs/1903.03894.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations, 2021. URL https://arxiv.org/abs/2102.05152.

A Table of Contents

B List of Symbols . . . 13
C Notation . . . 15
D Lie Groups and Lie Actions . . . 16
G Explanation Visualizations . . . 17
E Algorithm . . . 16
H Proofs of all Theorems and Propositions . . . 18
H.1 Proof of Theorem 1: Reductive Decomposition and Factored Group Action . . . 18
H.2 Proof of Proposition 1: Lie Algebra of the Label-Stabilizer . . . 18
H.3 Lemma 1 Codimension-one Structure of the Label Differential . . . 19
H.4 Proof of Theorem 2: Factual Orbits Foliate the Class Manifold . . . 19
F GLIDE Pipeline Diagram . . . 16
H.5 Proof of Theorem 3: Manifold Proximity Bound . . . 20
H.6 Proof of Theorem 4: Winner-Take-All Recovers the Minimum-Norm Generator . . . 20
H.7 Theorem 5: Proximity Bound Tightens Monotonically Over Training . . . 21
H.8 Theorem 6 with Proof: Two-Phase Ordering is Spectrally Necessary . . . 22
I Factual Orbit Learning: Orthogonality and Proximity Loss . . . 23
J Counterfactual Learning Losses . . . 24
J.1 Orthogonality Loss and Independence of Counterfactual Directions . . . 24
J.2 Adjoint Loss and Adjoint Consistency of the Decomposition . . . 25
J.3 Sparsity Loss . . . 25
J.3.1 Top-1 Generator Sparsity . . . 26
J.4 Margin Loss . . . 26
J.5 Proximity Loss . . . 26
K Factual and Counterfactual Loss Ablations . . . 26
L Symmetry Discovery via Lie Algebra Parameterization . . . 27
L.1 LieGAN and Augerino$^+$ as Factual Explainers . . . 27
L.2.1 LieGAN . . . 28
L.2.2 Augerino$^+$ . . . 28
L.2.3 Unified Evaluation . . . 28
L.2.4 Key Distinction from GLIDE . . . 28
L.3 Degeneracy of LieGAN and Augerino$^+$ Baselines . . . 29

M Counterfactual Baseline Details . . . 29
M.1 Shared Metrics . . . 29
M.2 Shared Adaptation . . . 30
M.3 CLEAR . . . 30
M.4 GCFExplainer . . . 30
M.5 D4Explainer . . . 30
M.6 Key Distinction from GLIDE . . . 30
M.7 Additional Ablations . . . 31
N Empirical Validation Details . . . 31
N.1 Validation of Proposition 1 Assumptions . . . 31
N.2 Validation of Theorem 2: Full Boundary Analysis . . . 32
N.3 Validation of Theorem 3: Manifold Proximity Guarantees . . . 33
N.4 Theorem 6: Spectral Validation Details . . . 33
N.5 Theorem 4: Generator Diagnostic Details . . . 34
O Linearized Gradient Flow . . . 36
P Extending GLIDE to Multi-Class Setting . . . 38
Q Guarantees and Bounds on Continuity and Differentiability . . . 38
R Computer Resources . . . 39

B List of Symbols

Symbol	Description
Graphs and Input Space	
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$	Attributed input graph
$\mathcal{V}$	Set of N nodes
$\mathcal{E}$	Set of edges
$X \in \mathbb{R}^{N \times d_0}$	Node feature matrix
N	Number of nodes
d_0	Node feature dimension
A_G	Adjacency matrix of graph $\mathcal{G}$
$\mathcal{G}_{\text{Sub}}$	Explanatory subgraph
Networks, Encoders, and Latent Space	
$f_\theta : \mathcal{G} \to \mathcal{Z}$	GNN encoder with parameters θ
$h_\phi : \mathcal{Z} \to \mathbb{R}^C$	Logit classifier with parameters ϕ
$g_\psi : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}^K$	Coefficient network with parameters ψ
$\mathcal{Z} \subset \mathbb{R}^D$	Latent embedding space
D	Latent space dimension
C	Number of classes
Δ^{C-1}	Probability simplex over C classes
$z = f_\theta(\mathcal{G})$	Graph embedding
z_f	Factually transformed embedding
z_{cf}, z'	Counterfactually transformed embedding
z^*	Minimal-displacement counterfactual embedding

Continued on next page

Symbol	Description
$\hat{y}(z) = \text{argmax}_c [h_\phi(z)]_c$	Predicted class label
y	Ground-truth class label
y_{cf}	Target counterfactual class label
Lie Groups, Lie Algebras, and Geometry	
G	Lie group acting smoothly on $\mathcal{Z}$
$e \in G$	Identity element of G
$\mathfrak{g} = T_e G$	Lie algebra of G (tangent space at e)
$[\cdot, \cdot]$	Lie bracket; for matrix groups, $[A, B] = AB - BA$
$\exp : \mathfrak{g} \to G$	Exponential map
$\text{Ad}_g(X) = gXg^{-1}$	Adjoint representation of G on $\mathfrak{g}$
$\text{ad}_X(Y) = [X, Y]$	Infinitesimal adjoint (adjoint action of $\mathfrak{g}$)
H_z	Label-stabilizer subgroup at z
H_z^0	Connected component of H_z containing e
$\mathfrak{h}_z = T_e H_z^0$	Lie algebra of H_z^0; factual subalgebra
$\mathfrak{m}_z = \mathfrak{h}_z^\perp$	Frobenius-orthogonal complement of $\mathfrak{h}_z$; counterfactual subspace
G/H_z^0	Reductive homogeneous space of counterfactual directions
$\mathcal{Z}_y$	Class manifold $\{z \in \mathcal{Z} : \hat{y}(z) = y\}$
$\partial \mathcal{Z}_y$	Decision boundary of class y
$\mathcal{O}_{\hat{y}}(z)$	Factual stabilizer orbit of z under class $\hat{y}$
$\mathcal{T}_{y_{\text{cf}}}(z)$	Counterfactual transverse orbit of z
$\langle A, B \rangle_F = \text{Tr}(A^\top B)$	Frobenius inner product
$\| \cdot \|_F$	Frobenius norm
$\ell_z(A) = \nabla_z y(z)^\top Az$	Linear functional defining $\mathfrak{h}_z$
Generators and Coefficients	
K	Number of Lie algebra basis generators
$\{A_k\}_{k=1}^K$	Learned Lie algebra basis generators, $A_k \in \mathbb{R}^{D \times D}$
$\{A_{f,k}\}_{k=1}^K$	Factual generators (label-preserving, Phase 1)
$\{A_{cf,k}\}_{k=1}^K$	Counterfactual generators (label-altering, Phase 2)
A_{f,k^*}	Dominant factual generator
A_{cf,k^*}	Dominant counterfactual generator
$k^* = \arg\max_k s_k$	Index of dominant generator
$U_k, V_k \in \mathbb{R}^{D \times r}$	Low-rank factors of generator $A_k = U_k V_k^\top$
r	Rank of low-rank generator factorization ($r \ll D$)
$t \in \mathbb{R}^K$	Coefficient vector for group transformation
$t_{cf} \in \mathbb{R}^K$	Counterfactual coefficient vector predicted by g_ψ
α_k	Randomly sampled factual coefficients (Phase 1 training)
$B = \sum_k t_{cf,k} A_{cf,k}$	Combined counterfactual transformation matrix
$s_k = \|\nabla_X t_{f,k}\|_2$	Saliency score for generator k over input X
$M = \max_k \|A_k\|_F$	Upper bound on generator Frobenius norms
α^*	Minimal fraction of B required to reach $\partial \mathcal{Z}_y$

Continued on next page

Symbol	Description
Loss Functions and Training	
$\mathcal{L}_{\text{ce}}$	Cross-entropy classification loss
$\mathcal{L}_{\text{total}}$	Total counterfactual training objective (Phase 2)
$\mathcal{L}_{\text{margin}}$	Margin-based label-flip loss
$\mathcal{L}_{\text{prox}}$	Manifold proximity loss ($\|z_{cf} - z\|_2^2$)
$\mathcal{L}_{\text{entropy}}$	Entropy-based sparsity loss (winner-take-all)
$\mathcal{L}_{\text{ortho}}$	Counterfactual generator orthogonality loss
$\mathcal{L}_{\text{adj}}$	Adjoint regularization loss (enforces reductive structure)
$\mathcal{L}^f_{\text{ortho}}$	Factual generator orthogonality loss (Phase 1)
$\mathcal{L}^f_{\text{prox}}$	Factual proximity loss (Phase 1)
$\lambda_1, \ldots, \lambda_5$	Loss weighting hyperparameters (Phase 2 objective)
γ	Decision margin hyperparameter
τ	Temperature parameter for Softmax sparsity
η	Gradient flow learning rate
Spectral and Hessian Quantities	
$H = \nabla^2 \mathcal{L}(\theta^*)$	Hessian of total training loss at critical point θ^*
$H\|_{\mathfrak{h}_z}$	Hessian restricted to factual subspace
$H\|_{\mathfrak{m}_z}$	Hessian restricted to counterfactual subspace
$\lambda_{\min}(H\|_{\mathfrak{h}_z})$	Minimum eigenvalue of factual Hessian
$\lambda_{\min}(H\|_{\mathfrak{m}_z})$	Minimum eigenvalue of counterfactual Hessian
$\lambda_* = \lambda_{\min}(H\|_{\mathfrak{m}_z})$	Minimum curvature of counterfactual subspace
$\Delta = \lambda_{\min}(H\|_{\mathfrak{h}_z}) - \lambda_{\min}(H\|_{\mathfrak{m}_z})$	Spectral gap between factual and counterfactual subspaces
$\tau_f = 1/\lambda_{\min}(H\|_{\mathfrak{h}_z})$	Characteristic convergence timescale of factual generators
$\tau_{cf} = 1/\lambda_{\min}(H\|_{\mathfrak{m}_z})$	Characteristic convergence timescale of counterfactual generators
$M(n) = \max_k \|A^{(n)}_{cf,k}\|_F$	Maximum counterfactual generator norm at gradient step n
Evaluation Metrics	
Fid^+	Necessity fidelity: prediction drop when dominant generator is removed
Fid^-	Sufficiency fidelity: target-class probability using only dominant generator
CF-Fid	Counterfactual fidelity: fraction of valid counterfactuals produced
CF-Prox	Counterfactual proximity: mean latent displacement $\|z_{cf} - z\|_2$
CF-Spar	Counterfactual sparsity: fraction of edges not perturbed
δ	Proximity tolerance ($\|z' - z\|_2 \leq \delta$)

C NOTATION

Let an attributed graph be $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$, where $\mathcal{V}$ is a set of N nodes, $\mathcal{E}$ is a set of edges, and $X \in \mathbb{R}^{N \times d_0}$ is the node feature matrix. A GNN encoder $f_\theta : \mathcal{G} \to \mathcal{Z} \subset \mathbb{R}^D$ produces a global graph embedding $z = f_\theta(\mathcal{G})$ via L rounds of neighborhood aggregation followed by a readout

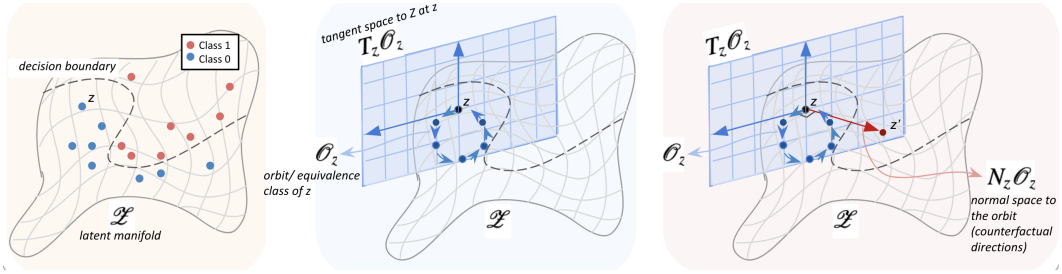

Figure 4: GLIDE latent space. Left Panel (Latent Manifold): the continuous latent space $\mathcal{Z}$ populated by graph embeddings, of Class 0 (blue circles) and Class 1 (red circles) separated by a dashed decision boundary. Middle Panel (Tangent Space & Factual Orbits): the tangent space $T_z\mathcal{O}_z$ at a given latent point z, with label-preserving factual flows along the orbit $\mathcal{O}_z$ that stay within the equivalence class. Right Panel (Counterfactual Direction): the normal space $N_z\mathcal{O}_z$, showing how counterfactual trajectories transition from z to a target embedding z' across the decision boundary.

function. A logit classifier $h_\phi : \mathcal{Z} \to \mathbb{R}^C$ produces a vector of class logits; the predicted label is $\hat{y}(z) = \operatorname{argmax}_c[h_\phi(z)]_c$, with probabilities $\operatorname{softmax}(h_\phi(z)) \in \Delta^{C-1}$. Throughout, we treat $\mathcal{Z}$ as a smooth Riemannian manifold with a non-trivial topology inherited from the discrete graph structure.

D Lie Groups and Lie Actions

A Lie group G is a smooth manifold that simultaneously carries a group structure compatible with its topology. The Lie algebra $\mathfrak{g} = T_eG$ is the tangent space at the identity element $e \in G$; it is a vector space closed under the Lie bracket $[\cdot, \cdot]$. For matrix groups, $[A, B] = AB - BA$.

The exponential map $\exp : \mathfrak{g} \to G$, defined by

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}, \tag{9}$$

is a local diffeomorphism near $0 \in \mathfrak{g}$. It provides a parametric family of smooth, invertible group elements: for any $A \in \mathfrak{g}$ and $t \in \mathbb{R}$, the curve $\gamma(t) = \exp(tA) \cdot z$ is a smooth one-parameter subgroup trajectory in G passing through z at $t = 0$.

The adjoint representation $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$ is defined by $\mathrm{Ad}_g(X) = gXg^{-1}$, with infinitesimal generator $\mathrm{Ad}_X(Y) = [X, Y]$. It characterizes how the group acts on its own Lie algebra and is central to our disentanglement construction.

E Algorithm

Our symmetry-explainable two-phase training protocol follows the algorithmic steps outlined in Section 5, and is detailed below. $\mathcal{L}_{\text{total}}$ is given in full in Section 5.2.

F GLIDE Pipeline Diagram

Figures 1 and 2 summarize the main architecture and geometric mechanics of the GLIDE framework. Figure 1 illustrates the foundational Lie-algebraic geometry of the latent space $\mathcal{Z}$, detailing how graph embeddings are decomposed into label-preserving factual stabilizer flows ($\mathfrak{h}_z$) and decision-crossing counterfactual trajectories ($\mathfrak{m}_z$). Figure 2 outlines the complete end-to-end execution pipeline—from frozen representation encoding and factual/ counterfactual generator optimization to the final input-space back-projection that yields graph saliency maps for factual explanations.

Algorithm 1 Symmetry-Explainable Training Protocol

Input: Graph dataset $\mathcal{D}$, encoder f_θ, factual generators $\{A_{f,k}\}$, CF generators $\{A_{cf,k}\}$, coefficient network g_ψ
Phase 1: Factual Orbit Learning
while not converged **do**
  Sample $(G, y) \sim \mathcal{D}$; compute $z \leftarrow f_\theta(G)$
  Sample $\alpha \sim \mathcal{N}(0, I_K)$
  $z_f \leftarrow \exp(\sum_k \alpha_k A_{f,k})\, z$
  Compute decision gradient $\nabla_z y(z)$
  Update $\{A_{f,k}\}$ to minimize

$$\mathcal{L}_{\text{ce}}(\text{softmax}(h_\phi(z_f)), y) + \lambda_{\text{tan}}\mathcal{L}_{\text{tan}} + \lambda_{\text{closure}}\mathcal{L}_{\text{closure}} + \lambda^f_{\text{ortho}}\mathcal{L}^f_{\text{ortho}} + \lambda^f_{\text{prox}}\mathcal{L}^f_{\text{prox}}$$

  { $\mathcal{L}_{\text{tan}}$: enforces $Az \in \ker \nabla_z y(z)$ (Prop. 1(ii))
  $\mathcal{L}_{\text{closure}}$: enforces approximate Lie closure $[\mathfrak{h}_z, \mathfrak{h}_z] \subset \mathfrak{h}_z$ }
end while
Freeze $\{A_{f,k}\}$
Phase 2: Counterfactual Reasoning
while not converged **do**
  Sample $(G, y) \sim \mathcal{D}$; $z \leftarrow f_\theta(G)$
  Sample target $y_{cf} \neq y$
  $t_{cf} \leftarrow g_\psi(z,\, y_{cf})$
  if epoch > 40 **then**
    $t_{cf} \leftarrow \text{HardMask}(t_{cf})$ {enforces 1D complement $\mathfrak{m}_z$}
  end if
  $z_{cf} \leftarrow \exp(\sum_k t_{cf,k} A_{cf,k})\, z$
  Compute $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}^{cf}_{ce} + \lambda^{cf}_{\text{prox}}\|z_{cf} - z\|^2 + \lambda_{\text{entropy}}\mathcal{H}(t_{cf}) + \lambda_{\text{ortho}}\mathcal{L}^{cf}_{\text{ortho}} + \lambda_{\text{adj}}\mathcal{L}_{\text{adjoint}}$$

  Clip gradients to norm 1.0
  Update $\{A_{cf,k}\}, \psi$
end while=0

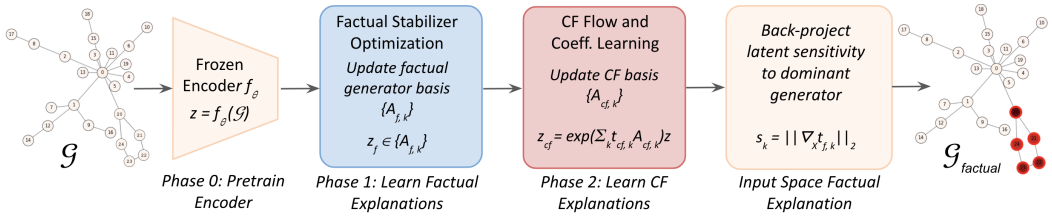

Figure 5: Phase 0 (Pretrain Encoder): The input graph $\mathcal{G}$ is passed through a frozen GNN encoder f_θ to yield the initial latent representation $z = f_\theta(\mathcal{G})$. Phase 1 (Learn Factual Explanations): Optimizes the factual generator basis $\{A_{\text{f},k}\}$ to model label-stabilizing transformations ($z_f \in \{A_{\text{f},k}\}$). Phase 2 (Learn CF Explanations): Computes counterfactual flows and coefficient learning via exponential mapping ($z_{\text{cf}} = \exp(\sum_k t_{\text{cf},k} A_{\text{cf},k}) z$) to update the counterfactual basis $\{A_{\text{cf},k}\}$. Input Space Factual Explanation: Back-projects latent sensitivities to the dominant generator to produce a highlighted subgraph saliency map ($\mathcal{G}_{\text{factual}}$).

G Explanation Visualizations

The saliency equations defining s_k in Definitions 2 and 3 back-project the dominant generator's direction onto node/edge-feature gradients (a soft, continuous saliency map), but not a discrete edit list like CF baselines, since GLIDE never decodes through a generative model. This is intrinsic, not an oversight. The latent-space operation is precisely what allows for the hard proximity bound (Theorem 3) and clean minimality (Theorem 4). Pairing GLIDE with a decoder remains open work.

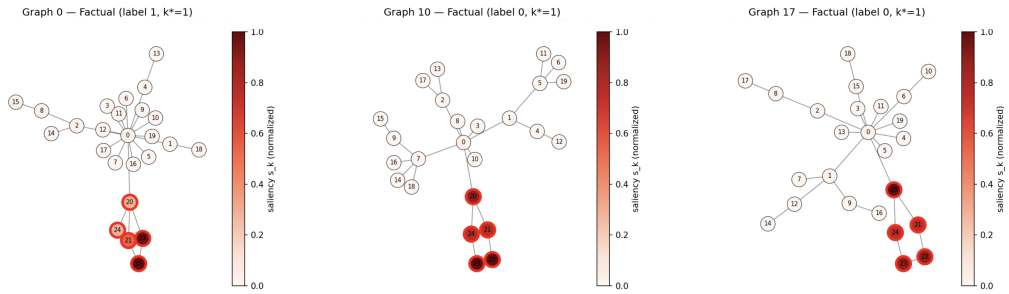

Figure 6: Visualizations of factual explanations on BA-2Motifs via GLIDE. GLIDE fully recovers the exact ground truth motifs on both house-motif graphs and cycle-motif graphs.

We have visualized the top-k nodes under the dominant factual generator to give direct evidence that the saliency-selected important nodes are correct. For BA-2Motifs, the top-5 nodes selected by the dominant factual generator overlap with the ground-truth motif exactly on 100% (for cycle) and 92% (for house) of test samples. All explanations that do not have full motif recovery identify ≥ 3 motif nodes. Full visualizations are included in Figure 6.

H Proofs of all Theorems and Propositions

In this section, we provide proofs of all theorems and propositions presented in the main paper, as well as the full statement and proof of Theorem H.7.

H.1 Proof of Theorem 1: Reductive Decomposition and Factored Group Action

We present the proof for Theorem 1 that was stated in Section 3.

Proof of Theorem 1. Since $\mathfrak{h}_z$ is a Lie subalgebra by Proposition 1 and the first condition of equation 5, it integrates to a Lie subgroup $H_z^0 \leq G$ (Lie's third theorem). The second condition of equation 5 states that $\mathfrak{m}_z$ is stable under the adjoint action $\mathrm{Ad}(H_z^0)$, i.e., $\mathrm{Ad}_h(\mathfrak{m}_z) \subseteq \mathfrak{m}_z$ for all $h \in H_z^0$. This is precisely the reductive decomposition condition (Etingof, 2026, Ch. 7, Prop. 7.1). The exponential map $\exp : \mathfrak{g} \to G$ is a local diffeomorphism near 0 (inverse function theorem on Lie groups). The map $\Phi : \mathfrak{h}_z \times \mathfrak{m}_z \to G$, $(X_{\mathrm{fact}}, X_{\mathrm{cf}}) \mapsto \exp(X_{\mathrm{fact}}) \exp(X_{\mathrm{cf}})$, has derivative at $(0, 0)$ equal to the identity on $\mathfrak{g}$, so by the inverse function theorem Φ is a local diffeomorphism. Hence, every $g \in G$ near e decomposes uniquely as in equation 6. □

H.2 Proof of Proposition 1: Lie Algebra of the Label-Stabilizer

We present the proof for Proposition 1 that was stated in Section 3.

Assumptions. We restate the three assumptions made in Proposition 1: Suppose the following three conditions hold at z:

1. **(Non-degeneracy)** $\nabla_z y(z) \neq 0$.
2. **(Faithfulness)** The evaluation map $\rho_z : \mathfrak{g} \to \mathbb{R}^D$, $A \mapsto Az$, is injective.
3. **(Local transitivity)** $\rho_z(\mathfrak{h}_z) = T_z\mathcal{Z}_y$, where $T_z\mathcal{Z}_y = \ker\big(\nabla_z y(z)^\top\big)$.

Proof. (i) Since $H_z^0 \leq G$ is a Lie subgroup by definition, its tangent space $\mathfrak{h}_z = T_e H_z^0$ is a Lie subalgebra of $\mathfrak{g}$ by standard Lie group theory.

(ii) We prove both inclusions separately.

($\mathfrak{h}_z \subseteq \ker \ell_z$): Let $A \in \mathfrak{h}_z$. The curve $\gamma(t) = \exp(tA) \cdot z$ lies in $H_z^0 \cdot z \subseteq \mathcal{Z}_y$ for all sufficiently small t, so $y(\gamma(t))$ is constant to first order. By the chain rule,

$$\ell_z(A) = \frac{d}{dt} y(\exp(tA) \cdot z)\Big|_{t=0} = \nabla_z y(z)^\top (Az) = 0, \tag{10}$$

hence $A \in \ker \ell_z$.

($\ker \ell_z \subseteq \mathfrak{h}_z$): Let $A \in \ker \ell_z$, so $\nabla_z y(z)^\top (Az) = 0$, i.e.,

$$Az \;\in\; \ker\big(\nabla_z y(z)^\top\big) \;=\; T_z \mathcal{Z}_y. \tag{11}$$

By local transitivity (Assumption 3), there exists $B \in \mathfrak{h}_z$ such that $Bz = Az$, i.e., $(A - B)z = 0$. By faithfulness (Assumption 2), ρ_z is injective, so $A - B = 0$ and hence $A = B \in \mathfrak{h}_z$.

Combining both inclusions: $\mathfrak{h}_z = \ker \ell_z$.

(iii) Since ρ_z is injective (Assumption 2) and $\nabla_z y(z) \neq 0$ (Assumption 1), the functional $\ell_z = \nabla_z y(z)^\top \circ \rho_z$ is a nonzero linear functional on $\mathfrak{g}$. To see that $\ell_z \not\equiv 0$: since $\nabla_z y(z) \neq 0$, there exists $v \in \mathbb{R}^D$ with $\nabla_z y(z)^\top v \neq 0$; injectivity of ρ_z then guarantees some $A_0 \in \mathfrak{g}$ with $A_0 z = v$, giving $\ell_z(A_0) \neq 0$. Applying the rank-nullity theorem to $\ell_z : \mathfrak{g} \to \mathbb{R}$:

$$\dim \ker \ell_z \;=\; \dim \mathfrak{g} - 1. \tag{12}$$

By part *(ii)*, $\dim \mathfrak{h}_z = \dim \mathfrak{g} - 1$. □

H.3 Lemma 1 Codimension-one Structure of the Label Differential

Assumption 1 (Non-degenerate group action)**.** *There exists $A \in \mathfrak{g}$ such that $\nabla_z y(z)^\top (Az) \neq 0$.*

Lemma 1 (Codimension-one structure of the label differential)**.** *Let $\ell_z(A) = \nabla_z y(z)^\top (Az)$. If $\nabla_z y(z) \neq 0$ and the infinitesimal action is non-trivial, then ℓ_z is a nonzero linear functional on $\mathfrak{g}$. Therefore,*

$$\dim \ker \ell_z = \dim \mathfrak{g} - 1,$$

so $\ker \ell_z$ is a hyperplane in $\mathfrak{g}$.

Proof. The map $A \mapsto Az$ is linear, and $\nabla_z y(z)^\top (\cdot)$ is a linear functional on $\mathbb{R}^d$. Hence ℓ_z is linear on $\mathfrak{g}$. By the non-degeneracy assumption (Assumption 1) that the infinitesimal action is not contained in $\ker(\nabla_z y(z)^\top)$, there exists $A_0 \in \mathfrak{g}$ such that

$$\nabla_z y(z)^\top (A_0 z) \neq 0,$$

so ℓ_z is not identically zero.

Since $\mathfrak{g}$ is finite-dimensional, the rank-nullity theorem implies

$$\dim \ker \ell_z = \dim \mathfrak{g} - 1.$$

□

H.4 Proof of Theorem 2: Factual Orbits Foliate the Class Manifold

We present the proof for Theorem 2 that was stated in Section 3.

Proof of Theorem 2. **Foliation.** Since $\mathfrak{h}_z$ is a Lie subalgebra (Proposition 1), it integrates to a Lie subgroup $H_z^0 \leq G$ (Lie's third theorem). The orbits $\mathcal{O}_y(z) = H_z^0 \cdot z$ are the leaves of the orbit foliation induced by the action of H_z^0 on $\mathcal{Z}_y$. By the Frobenius theorem, a smooth distribution on a manifold that is closed under the Lie bracket defines a foliation; since $\mathfrak{h}_z$ is closed under the bracket by Proposition 1, the corresponding distribution $z \mapsto \{A_{f,k} z\}$ is its own inverse, and the Frobenius theorem guarantees that it integrates to a smooth foliation of $\mathcal{Z}_y$.

i. Completeness

Fix $z \in \mathcal{Z}_y$. Since $\mathfrak{h}_z$ integrates to a Lie subgroup H_z^0 (by Lie's third theorem), the map

$$X_{\text{fact}} \mapsto \exp(X_{\text{fact}}) \cdot z$$

defines a smooth local chart from a neighborhood of $0 \in \mathfrak{h}_z$ onto the orbit $\mathcal{O}_y(z) = H_z^0 \cdot z$.

By Proposition 1, this orbit is an immersed submanifold tangent to the stabilizer distribution and contained in $\mathcal{Z}_y$. Hence it coincides locally with the leaf of the stabilizer foliation through z.

Therefore, for any z' lying in the same local leaf as z (i.e., in the same connected component of the stabilizer orbit inside a sufficiently small neighborhood), there exists $X_{\text{fact}} \in \mathfrak{h}_z$ such that

$$z' = \exp(X_{\text{fact}}) \cdot z.$$

Thus the factual generators are complete in the sense that they parameterize all local variations of z within its stabilizer leaf.

ii. Decision boundary The decision boundary $\partial\mathcal{Z}_y$ is by definition the set of points z such that every neighborhood of z contains points in both $\mathcal{Z}_y$ and $\mathcal{Z}_{y'}$ for some $y' \neq y$. A point $z^* \in \partial\mathcal{Z}_y$ satisfies $y(z^*) = y$ marginally, meaning $\nabla_z y(z^*)^\top X_{\text{cf}} z^* > 0$ for some $X_{\text{cf}} \in \mathfrak{m}_z$. The orbit $\mathcal{T}_{y_{\text{cf}}}(z)$ terminates at the first t^* for which $y(\exp(t^* A_{cf,k}) \cdot z) = y_{\text{cf}}$; this first-crossing point lies on $\partial\mathcal{Z}_y$. Since every $z^* \in \partial\mathcal{Z}_y$ is reachable as such a first-crossing point from some interior z, the union of transverse orbit leaves covers $\partial\mathcal{Z}_y$. □

H.5 Proof of Theorem 3: Manifold proximity Bound

We provide the proof for Theorem 3 that was stated in Section 3.

Proof of Theorem 3. Let $B = \sum_{k=1}^K t_k A_k$. By the triangle inequality and sub-multiplicativity of the operator norm,

$$\|B\|_F \leq \sum_{k=1}^K |t_k| \|A_k\|_F \leq M \|t\|_1. \tag{13}$$

For any matrix B, the matrix exponential satisfies

$$\|\exp(B) - I\|_F \leq e^{\|B\|_F} - 1, \tag{14}$$

which follows from the triangle inequality applied to the power series $\sum_{n=1}^\infty \frac{B^n}{n!}$: $\|\exp(B) - I\|_F \leq \sum_{n=1}^\infty \frac{\|B\|_F^n}{n!} = e^{\|B\|_F} - 1$. Therefore,

$$\|z' - z\|_2 = \|(\exp(B) - I)z\|_2 \leq \|\exp(B) - I\|_F \|z\|_2 \leq \|z\|_2 \left(e^{\|B\|_F} - 1\right) \leq \|z\|_2 \left(e^{M\|t\|_1} - 1\right). \tag{15}$$

The bound on $\|t\|_1$ for proximity δ follows by inverting the inequality. □

H.6 Proof of Theorem 4: Winner-Take-All Recovers the Minimum-Norm Generator

We provide the proof for Theorem 4 that was stated in Section 5.4.

Proof of Theorem 4. **Step 1: Scalar reduction.** By Proposition 1(ii)–(iii), $\mathfrak{m}_z$ is one-dimensional and directions in $\mathfrak{h}_z$ contribute zero to the classifier output ℓ_z. Hence the first-order label-flip condition depends only on the projection of $\sum_k t_k A_{cf,k}$ onto m, reducing to the scalar constraint

$$\left|\sum_{k=1}^K t_k a_k\right| \geq \theta. \tag{16}$$

Step 2: Architecture induces one-hot selection. As $\tau \to 0$, minimizing $\mathcal{L}_{\text{entropy}}$ drives $\text{Softmax}(t_{cf}/\tau)$ toward a point mass. The hard Gumbel-Softmax mask then projects the solution onto the corresponding one-hot vector $t = s\, e_k$ for some $k \in [K]$ and scalar $s > 0$.

Step 3: Minimization over one-hot vectors. For $t = s\, e_k$, constraint equation 16 becomes $|s||a_k| \geq \theta$, attained at equality with $|s_k^*| = \theta/|a_k|$. The corresponding latent displacement satisfies

$$\|\Delta z\|_2 \;\approx\; |s_k^*|\, \|A_{cf,k} z\|_2 \;=\; \frac{\theta\, \|A_{cf,k} z\|_2}{|a_k|}. \tag{17}$$

Minimizing over k gives

$$k^* = \arg\max_k \frac{|a_k|}{\|A_{cf,k} z\|_2}. \tag{18}$$

This maximizer is unique under non-degenerate projections $\{a_k\}$. □

H.7 Theorem 5: Proximity Bound Tightens Monotonically Over Training

We include Theorem 5 to strengthen the claims of Theorem 3, and prove the bounded and monotonically decreasing proximity of counterfactual explanations to the original graphs over training.

Theorem 5 (Proximity Bound Tightens Monotonically Over Training)**.** *Let $A_k^* = \lim_{n\to\infty} A_{cf,k}^{(n)}$ denote the fixed point of the combined training loss $\mathcal{L}_{\text{total}}$, which balances $\mathcal{L}_{\text{prox}}$ (pushing generators toward zero) and $\mathcal{L}_{\text{margin}}$ (holding generators away from zero). Define*

$$M^* = \max_k \|A_k^*\|_F, \qquad \delta A_k^{(n)} := A_{cf,k}^{(n)} - A_k^*. \tag{19}$$

Let $\lambda^* = \lambda_{\min}(H|_{\mathcal{S}_{cf}}) > 0$ denote the minimum curvature of $\mathcal{L}_{\text{total}}$ restricted to the counterfactual subspace $\mathcal{S}_{cf} = \text{span}\{A_{cf,k}^{(0)}\}$ at the fixed point A_k^*. Under gradient descent with step size $\eta > 0$, the deviation from the fixed point decays as:

$$\|\delta A_k^{(n)}\|_F \;\le\; e^{-\eta\lambda^* n}\, \|\delta A_k^{(0)}\|_F, \tag{20}$$

and consequently the generator norms satisfy:

$$\|A_{cf,k}^{(n)}\|_F \;\le\; M^* + e^{-\eta\lambda^* n}\, \|\delta A_k^{(0)}\|_F. \tag{21}$$

The counterfactual proximity bound of Theorem 3 therefore tightens monotonically from its initial value to the fixed-point bound:

$$\|z' - z\|_2 \;\le\; \|z\|_2 \Big(\exp\Big[\Big(M^* + e^{-\eta\lambda^* n}\|\delta A^{(0)}\|\Big)\,\|t\|_1\Big] - 1\Big) \;\xrightarrow{n\to\infty}\; \|z\|_2 \Big(e^{M^*\|t\|_1} - 1\Big), \tag{22}$$

where $\|\delta A^{(0)}\| = \max_k \|\delta A_k^{(0)}\|_F$. The bound is monotonically non-increasing in n and converges to the tightest achievable value $\|z\|_2(e^{M^*\|t\|_1} - 1)$, determined by the fixed point of $\mathcal{L}_{\text{total}}$ rather than by $\mathcal{L}_{\text{prox}}$ alone.

Proof of Theorem 5. **Step 1: Fixed point and linearization.** The combined loss $\mathcal{L}_{\text{total}}$ includes $\mathcal{L}_{\text{prox}} = \|z_{cf} - z\|_2^2$ and $\mathcal{L}_{\text{margin}} = \mathbb{E}[\max(0, \gamma - (\text{logit}_{y_{cf}} - \text{logit}_y))]$, which exert opposing gradients on $A_{cf,k}$. Their balance determines a non-zero fixed point A_k^* satisfying $\nabla_{A_{cf,k}} \mathcal{L}_{\text{total}}|_{A_k^*} = 0$. By a first-order Taylor expansion of the gradient near A_k^*:

$$\nabla_{A_{cf,k}} \mathcal{L}_{\text{total}} = H\, \delta A_k^{(n)} + o\Big(\|\delta A_k^{(n)}\|\Big), \tag{23}$$

where $H = \nabla^2 \mathcal{L}_{\text{total}}|_{A_k^*}$ is the Hessian at the fixed point. This linearization is valid for $\|\delta A_k^{(n)}\|_F$ sufficiently small, i.e., in a neighborhood of convergence.

Step 2: Exponential contraction of deviations. Restricting to $\mathcal{S}_{cf}$ and using $H|_{\mathcal{S}_{cf}} \succeq \lambda^* I$ with $\lambda^* > 0$, the gradient descent update $A_{cf,k}^{(n+1)} = A_{cf,k}^{(n)} - \eta \nabla_{A_{cf,k}} \mathcal{L}_{\text{total}}$ gives:

$$\delta A_k^{(n+1)} \approx (I - \eta H)\, \delta A_k^{(n)}, \quad \|I - \eta H\| \le e^{-\eta\lambda^*}, \tag{24}$$

hence $\|\delta A_k^{(n)}\|_F \le e^{-\eta\lambda^* n} \|\delta A_k^{(0)}\|_F$, establishing equation 20.

Step 3: Bound on generator norms. By the triangle inequality,

$$\|A_{cf,k}^{(n)}\|_F = \|A_k^* + \delta A_k^{(n)}\|_F \le \|A_k^*\|_F + \|\delta A_k^{(n)}\|_F \le M^* + e^{-\eta\lambda^* n} \|\delta A_k^{(0)}\|_F, \tag{25}$$

establishing equation 21. As $n \to \infty$, $\|A_{cf,k}^{(n)}\|_F \to M^* = \|A_k^*\|_F$.

Step 4: Substitution into the proximity bound. Let $B^{(n)} = \sum_k t_k A^{(n)}_{cf,k}$. Then:

$$\|B^{(n)}\|_F \leq \sum_k |t_k| \|A^{(n)}_{cf,k}\|_F \leq \left(M^* + e^{-\eta\lambda^* n}\|\delta A^{(0)}\|\right) \|t\|_1, \tag{26}$$

and by the matrix exponential bound from Theorem 3:

$$\|z' - z\|_2 \leq \|z\|_2 \left(\exp\left[\left(M^* + e^{-\eta\lambda^* n}\|\delta A^{(0)}\|\right) \|t\|_1\right] - 1\right). \tag{27}$$

Step 5: Monotonicity and convergence. Since $e^{-\eta\lambda^* n}$ is strictly decreasing in n, the upper bound in Step 4 is strictly decreasing in n. As $n \to \infty$, it converges to $\|z\|_2(e^{M^*\|t\|_1} - 1)$, the tightest bound achievable under $\mathcal{L}_{\text{total}}$. This establishes equation 22. □

H.8 Theorem 6 with Proof: Two-Phase Ordering is Spectrally Necessary

Theorem 6 (Two-Phase Ordering is Spectrally Necessary)**.** *Let $H|_{\mathfrak{h}_z}$ and $H|_{\mathfrak{m}_z}$ denote the Hessian of the total training loss restricted to the factual and counterfactual subspaces respectively. Under the assumption that the GNN encoder is G-equivariant and the classification loss is G-invariant, the minimum curvatures satisfy:*

$$\lambda_{\min}(H|_{\mathfrak{h}_z}) \;\geq\; \lambda_{\min}(H|_{\mathfrak{m}_z}), \tag{28}$$

with strict inequality whenever the decision boundary is non-degenerate. The characteristic convergence timescales satisfy $\tau_f \leq \tau_{cf}$, so factual generators converge no slower than counterfactual generators under gradient flow. Furthermore, joint training of both generator sets introduces cross-subspace interference of magnitude:

$$\epsilon_{\text{interference}} \;\leq\; \frac{\|H_{hm}\|_F}{\lambda_{\min}(H|_{\mathfrak{h}_z})}, \qquad H_{hm} = \nabla^2_{A_f,\, A_{cf}} \mathcal{L}_{\text{total}}, \tag{29}$$

which vanishes when the subspaces are Hessian-orthogonal and is bounded away from zero in the joint training setup.

Preliminary: Irreducibility of $\mathfrak{h}_z$ and $\mathfrak{m}_z$. We first establish that $\mathfrak{h}_z$ and $\mathfrak{m}_z$ are irreducible as G-representations under the adjoint action $\mathrm{Ad} : G \to GL(\mathfrak{g})$, and that they are mutually non-isomorphic. This is required to apply Schur's lemma in Step 1 of the proof below.

Lemma 2 (Irreducibility and Non-Isomorphism)**.** *Under the non-degeneracy condition of Theorem 6:*

1. *$\mathfrak{m}_z$ is irreducible as a G-representation.*
2. *$\mathfrak{m}_z$ and $\mathfrak{h}_z$ are mutually non-isomorphic.*
3. *$\mathfrak{h}_z$ is irreducible under the Genericity Assumption (GA).*

Proof. (i) By Proposition 1(iii), $\mathfrak{h}_z$ has codimension one in $\mathfrak{g}$, so $\mathfrak{m}_z = \mathfrak{h}_z^{\perp}$ is one-dimensional. Every one-dimensional representation space is trivially irreducible.

(ii) Since $\dim \mathfrak{m}_z = 1$ and $\dim \mathfrak{h}_z = \dim \mathfrak{g} - 1 \geq 1$, the two subspaces have strictly different dimensions and therefore cannot be isomorphic as G-representations.

(iii) By the Genericity Assumption (GA) stated below.

Assumption 2 (Genericity Assumption (GA))**.** *The adjoint action of G on $\mathfrak{h}_z$ is irreducible, i.e., $\mathfrak{h}_z$ admits no proper nonzero $\mathrm{Ad}(G)$-invariant subspace.*

This holds whenever the label-stabilizer subgroup H^0_x is a simple Lie group, equivalently when $\mathfrak{h}_z$ is a simple Lie algebra. For generic GNN encoders with non-degenerate decision boundaries, the stabilizer subgroup does not decompose as a non-trivial direct product of smaller subgroups, so (GA) is generically satisfied. When (GA) fails, $H|_{\mathfrak{h}_z}$ further decomposes into irreducible blocks, and Theorem 6 applies within each component; the curvature ordering $\lambda_{\min}(H|_{\mathfrak{h}_z}) \geq \lambda_{\min}(H|_{\mathfrak{m}_z})$ and the conclusion $\tau_f \leq \tau_{cf}$ are unaffected. □

Step 1: Hessian block structure. The G-invariance of $\mathcal{L}$ implies $H\rho(g) = \rho(g)H$ for all $g \in G$, so H commutes with the G-action on $\mathfrak{g}$. By Lemma 2, $\mathfrak{g} = \mathfrak{h}_z \oplus \mathfrak{m}_z$ is a decomposition into non-isomorphic irreducible G-representations (under (GA)). By Schur's lemma, any G-equivariant endomorphism of a direct sum of non-isomorphic irreducibles is block-diagonal over that decomposition. Therefore:

$$H = \begin{pmatrix} H|_{\mathfrak{h}_z} & 0 \\ 0 & H|_{\mathfrak{m}_z} \end{pmatrix}. \tag{30}$$

The off-diagonal blocks vanish exactly because $\mathfrak{h}_z$ and $\mathfrak{m}_z$ carry distinct irreducible characters, so no G-equivariant map can send one to the other (Schur's lemma, non-isomorphic case).

Alternative argument (without (GA)). Even without (GA), block structure follows directly from loss additivity. The total loss separates as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fact}}(\{A_{f,k}\}) + \mathcal{L}_{cf}(\{A_{cf,k}\}) + \lambda_5 \mathcal{L}_{\text{adj}}$. In the limit $\mathcal{L}_{\text{adj}} \to 0$ (enforced by training), there are no cross-terms between the factual and counterfactual generator parameters, so H is exactly block-diagonal by additivity of second derivatives. This argument requires no representation-theoretic assumptions and holds whenever training has converged.

Step 2: Curvature ordering. The factual subspace $\mathfrak{h}_z$ spans label-preserving directions, along which the classification loss $\mathcal{L}_{ce}$ is flat. However, the full training objective includes the manifold proximity loss $\mathcal{L}_{\text{prox}}$ and orthogonality loss $\mathcal{L}_{\text{ortho}}$, which are strongly convex (positive definite Hessian) in the factual generator parameters. The factual generators are therefore trained against a positive-definite curvature contribution, giving $\lambda_{\min}(H|_{\mathfrak{h}_z}) \geq \lambda_p > 0$.

The counterfactual subspace $\mathfrak{m}_z$ spans directions that must cross the decision boundary. The margin loss $\mathcal{L}_{\text{margin}}$ provides curvature in these directions, but only near the boundary; far from the boundary it saturates and contributes zero curvature. The effective minimum curvature $\lambda_{\min}(H|_{\mathfrak{m}_z})$ is therefore controlled by the density of training samples near the decision boundary, which is generically smaller than λ_p. Strict inequality holds whenever the boundary has positive codimension in $\mathcal{Z}$, which holds whenever the classifier is non-degenerate.

Step 3: Convergence timescales. By the standard analysis of linearized gradient flow (Appendix O), the convergence timescale of subspace λ is $\tau_\lambda = 1/\lambda_{\min}(H_\lambda)$. The ordering $\lambda_{\min}(H|_{\mathfrak{h}_z}) \geq \lambda_{\min}(H|_{\mathfrak{m}_z})$ directly gives $\tau_f \leq \tau_{cf}$.

Step 4: Interference bound. Under joint training with finite $\mathcal{L}_{\text{adj}}$, the full Hessian acquires off-diagonal blocks:

$$H_{\text{joint}} = \begin{pmatrix} H|_{\mathfrak{h}_z} & H_{hm} \\ H_{mh} & H|_{\mathfrak{m}_z} \end{pmatrix}, \qquad H_{hm} = \nabla^2_{A_f,\, A_{cf}} \mathcal{L}_{\text{total}}, \tag{31}$$

where H_{hm} vanishes exactly when $\mathcal{L}_{\text{adj}} \to 0$ (established in Step 1). The gradient update for the factual generators under joint training receives the spurious cross-subspace contribution $H_{hm}\, \delta A_{cf}$. By sub-multiplicativity of the Frobenius norm and the curvature lower bound $\|H|_{\mathfrak{h}_z}\, \delta A_f\|_F \geq \lambda_{\min}(H|_{\mathfrak{h}_z}) \|\delta A_f\|_F$, the relative interference satisfies:

$$\epsilon_{\text{interference}} := \frac{\|H_{hm}\, \delta A_{cf}\|_F}{\|H|_{\mathfrak{h}_z}\, \delta A_f\|_F} \leq \frac{\|H_{hm}\|_F}{\lambda_{\min}(H|_{\mathfrak{h}_z})}. \tag{32}$$

This vanishes as $\mathcal{L}_{\text{adj}} \to 0$ and is bounded away from zero in joint training whenever $H_{hm} \neq 0$. □

I Factual Orbit Learning: Tangent Alignment and Locality

As introduced in Section 5.2, we incorporate losses that enforce that factual generators lie in the tangent space of the decision boundary and generate locally stable orbits.

Tangent alignment. To ensure that factual generators preserve the predicted label to first order, we enforce alignment with the decision boundary via

$$\mathcal{L}^f_{\text{tan}} = \sum_i \big(\nabla_z y(z)^\top (A_{f,i} z)\big)^2, \tag{33}$$

which penalizes motion in the normal direction and encourages each generator to lie in the kernel of the linearized decision functional ℓ_z.

Table 4: Loss terms in GLIDE's two-phase objective. Appendix K reports full ablation tables for all loss terms listed below.

Term	Eq.	Theory	Phase	Function
$\mathcal{L}_{ce}$	–	–	1	Cross-entropy pressure driving z_f into the correct class region
$\mathcal{L}^f_{tan}$	Eq. 41	Prop. 1(ii)	1	Penalizes normal-direction motion; keeps factual generators tangent to the level set
$\mathcal{L}^f_{closure}$	Eq. 42	Prop. 1(i)	1	Encourages bracket-closure, i.e. a genuine Lie subalgebra rather than arbitrary directions
$\mathcal{L}^f_{prox}$	Eq. 43	–	1	Keeps z_f near z; prevents trivial large/global transformations from satisfying $\mathcal{L}_{ce}$
$\mathcal{L}^f_{ortho}$	Eq. 44	Prop. 2 (factual)	1	Discourages redundant/degenerate factual generators
$\mathcal{L}_{margin}$	Eq. 49	Def. 3	2	Pushes z_{cf} across the boundary with confidence margin γ, not just barely across it
$\mathcal{L}_{prox}$	Eq. 50	Thm. 3	2	Directly bounds $\|z_{cf} - z\|^2$, the quantity Thm. 3's proximity guarantee is stated about
$\mathcal{L}_{entropy}$	Eq. 48	Thm. 4	2	Winner-take-all sparsity; drives t_{cf} toward one-hot, enabling minimal-generator selection
$\mathcal{L}_{ortho}$	Eq. 45	Prop. 2	2	Frobenius-orthogonality of CF generators; distinct generators induce distinct latent directions
$\mathcal{L}_{adj}$	Eq. 47	Prop. 3 / Thm. 1	2	Penalizes $[A_{f,i}, A_{cf,j}]$ falling outside $\mathfrak{m}_z$; drives the reductive pair $(\mathfrak{h}_z, \mathfrak{m}_z)$

Lie closure. To promote a coherent stabilizer structure, we encourage the generators to form an approximate Lie subalgebra:

$$\mathcal{L}^f_{\text{closure}} = \sum_{i,j} \left\| [A_{f,i}, A_{f,j}] - \text{proj}_{\text{span}\{A_{f,k}\}}([A_{f,i}, A_{f,j}]) \right\|_F^2 . \tag{34}$$

This enforces closure under the Lie bracket, ensuring that the learned directions are stable under composition.

Proximity. Finally, we enforce that factual transformations remain in a local neighborhood of the input:

$$\mathcal{L}^f_{\text{prox}} = \|z_f - z\|_2^2, \tag{35}$$

which prevents trivial global transformations and ensures the learned orbit approximates a local symmetry of the decision region.

Orthogonality. For numerical stability, we optionally include an orthogonality regularizer

$$\mathcal{L}^f_{\text{ortho}} = \sum_{i \neq j} \|A_{f,i}^\top A_{f,j}\|_F^2, \tag{36}$$

which discourages degenerate or redundant generators but is not required for the theoretical guarantees.

J Counterfactual Learning Losses

As introduced in Section 5.2, our total counterfactual objective includes an orthogonality loss, adjoint loss, sparsity loss and a margin-based label flip loss, as defined below.

J.1 Orthogonality Loss and Independence of Counterfactual Directions

We define

$$\mathcal{L}_{\text{ortho}} = \sum_{i \neq j} \|A_{cf,i}^\top A_{cf,j}\|_F^2. \tag{37}$$

For explanations to be concise and non-redundant, the counterfactual generators should induce locally independent directions of motion in $\mathcal{Z}$. Since orthogonality of generator matrices in $\mathfrak{g}$ does not imply orthogonality of their induced motions on a specific $z \in \mathcal{Z}$, we must establish the stronger condition under which distinct generators provably explore distinct explanatory directions on the data.

Proposition 2 (Orthogonal Generators Yield Locally Independent Directions)**.** *Suppose $\langle A_i, A_j \rangle_F = \mathrm{Tr}(A_i^\top A_j) = 0$ for all $i \neq j$ (Frobenius orthogonality). Then the latent displacement vectors $\{A_k z\}_{k=1}^K$ are mutually near-orthogonal in $\mathbb{R}^D$, with*

$$|\langle A_i z, A_j z \rangle| \leq \|A_i^\top A_j\|_F \|z\|_2^2 = 0. \tag{38}$$

In particular, the directions $\{A_k z / \|A_k z\|_2\}$ are linearly independent whenever $A_k z \neq 0$ for all k.

Proof of Proposition 2. By the Cauchy-Schwarz inequality for Frobenius inner products, $|\langle A_i z, A_j z \rangle| = |z^\top A_i^\top A_j z| \leq \|A_i^\top A_j\|_F \|z\|_2^2$. Under Frobenius orthogonality, $\mathrm{Tr}(A_i^\top A_j) = 0$; however, note that $A_i^\top A_j$ may not be zero even when $\mathrm{Tr}(A_i^\top A_j) = 0$. For a stronger independence guarantee, we enforce the stronger condition $\|A_i^\top A_j\|_F = 0$, i.e., $A_i^\top A_j = 0$, via the orthogonality loss $\mathcal{L}_{\mathrm{ortho}} = \sum_{i \neq j} \|A_i^\top A_j\|_F^2$. Under this condition, equation equation 38 gives $\langle A_i z, A_j z \rangle = 0$, establishing mutual orthogonality in $\mathcal{Z}$. Linear independence follows because a set of mutually orthogonal non-zero vectors is linearly independent. □

Proposition 2 justifies the orthogonality loss $\mathcal{L}_{\mathrm{ortho}}$ not merely as a regularization heuristic but as a condition that guarantees the generators explore structurally distinct explanatory directions. Without this, multiple generators could discover redundant perturbations, violating the conciseness requirement of counterfactual explanations.

J.2 Adjoint Loss and Adjoint Consistency of the Decomposition

The reductive condition $[\mathfrak{h}_z, \mathfrak{m}_z] \subseteq \mathfrak{m}_z$ (Theorem 1) is not automatically satisfied by the learned generators. We introduce an adjoint regularization loss that enforces it in Proposition 3.

Proposition 3 (Adjoint Regularization Enforces Reductive Structure)**.** *Define the adjoint inconsistency of the learned factual generators $\{A_{f,k}\}$ and counterfactual generators $\{A_{cf,k}\}$ as*

$$\mathcal{L}_{\mathrm{adj}} = \sum_{i,j} \left\| \mathrm{proj}_{\mathfrak{h}_z}([A_{f,i}, A_{cf,j}]) \right\|_F^2, \tag{39}$$

where $\mathrm{proj}_{\mathfrak{h}_z}$ denotes projection onto the span of $\{A_{f,k}\}$. Minimizing $\mathcal{L}_{\mathrm{adj}} \to 0$ is a necessary and sufficient condition for the learned decomposition to satisfy the reductive condition $[\mathfrak{h}_z, \mathfrak{m}_z] \subseteq \mathfrak{m}_z$.

Proof. The reductive condition requires $[A_{f,i}, A_{cf,j}] \in \mathfrak{m}_z$ for all i, j, equivalently $\mathrm{proj}_{\mathfrak{h}_z}([A_{f,i}, A_{cf,j}]) = 0$. Since $\|\cdot\|_F^2$ is a norm, this condition holds for all i, j if and only if the sum in equation 39 is zero. Minimizing $\mathcal{L}_{\mathrm{adj}}$ over the generator parameters therefore drives the decomposition toward the reductive structure required by Theorem 1, making the factorization equation 6 increasingly exact as training progresses. □

J.3 Sparsity Loss

To recover the minimal-norm counterfactual generator (Theorem 4), we apply a temperature-scaled Softmax to t_{cf} and minimize the entropy:

$$\mathcal{L}_{\mathrm{entropy}} = -\frac{1}{B} \sum_{b=1}^{B} \sum_{k=1}^{K} P(t_{cf}^{(k)}) \log P(t_{cf}^{(k)}), \tag{40}$$

where $P(t_{cf}^{(k)}) = \mathrm{Softmax}(t_{cf}/\tau)_k$. After epoch 40, we apply a hard Gumbel-Softmax mask, collapsing to the winning generator.

Table 5: Top-1 sparsity of counterfactual explanations from GLIDE. Results are mean ± std over 10 runs. Higher is better.

	BA_2Motifs	**MUTAG**	**PROTEINS**	**NCI1**	**AIDS**
Top-1 Sparsity (↑)	0.91 ± 0.08	0.88 ± 0.09	0.79 ± 0.29	0.75 ± 0.25	0.86 ± 0.33

Table 6: Fid- results across datasets and loss ablations. All loss terms are important. Full GLIDE loss has the best performance.

Dataset	Full GLIDE Loss	-$\mathcal{L}_{closure}$	-$\mathcal{L}_{ortho}$	-$\mathcal{L}_{prox}$	-$\mathcal{L}_{tan}$
AIDS	$\mathbf{0.95 \pm 0.01}$	0.93 ± 0.01	0.94 ± 0.00	0.92 ± 0.01	0.93 ± 0.01
BA_2Motifs	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
MUTAG	$\mathbf{0.98 \pm 0.10}$	0.94 ± 0.01	0.94 ± 0.01	0.93 ± 0.01	0.94 ± 0.01
NCI1	$\mathbf{0.78 \pm 0.04}$	0.75 ± 0.01	0.76 ± 0.00	0.75 ± 0.01	0.76 ± 0.00
PROTEINS	$\mathbf{0.83 \pm 0.02}$	0.71 ± 0.02	0.69 ± 0.02	0.71 ± 0.01	0.70 ± 0.03

Table 7: Fid+ results across datasets and loss ablations. All loss terms are important. Full GLIDE loss has the best performance.

Dataset	Full GLIDE Loss	-$\mathcal{L}_{closure}$	-$\mathcal{L}_{ortho}$	-$\mathcal{L}_{prox}$	-$\mathcal{L}_{tan}$
AIDS	0.05 ± 0.01	$\mathbf{0.01 \pm 0.01}$	$\mathbf{0.01 \pm 0.00}$	0.02 ± 0.00	$\mathbf{0.01 \pm 0.00}$
BA_2Motifs	$\mathbf{0.00 \pm 0.00}$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
MUTAG	$\mathbf{0.02 \pm 0.01}$	0.02 ± 0.00	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.00
NCI1	$\mathbf{0.00 \pm 0.00}$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
PROTEINS	$\mathbf{0.04 \pm 0.01}$	0.04 ± 0.01	0.05 ± 0.02	0.04 ± 0.02	0.05 ± 0.01

J.3.1 Top-1 Generator Sparsity

Generator selection statistics confirm that the winner-take-all mechanism converges to a dominant generator ($\geq 75\%$ selection rate for k^* across datasets), consistent with Theorem 4. This indicates that a single Lie algebra direction suffices to characterize the model's decision boundary crossing, providing an interpretable and concise counterfactual explanation.

J.4 Margin Loss

We use a margin-based label flip loss to ensure the counterfactual penetrates the decision boundary:

$$\mathcal{L}_{\text{margin}} = \mathbb{E}\Big[\max\Big(0,\ \gamma - (\text{logit}_{y_{cf}} - \text{logit}_{y})\Big)\Big], \tag{41}$$

where $\gamma > 0$ controls the required decision margin.

J.5 Proximity Loss

We define the proximity loss as

$$\mathcal{L}_{\text{prox}} = \|z_{cf} - z\|_2^2. \tag{42}$$

This enforces proximity on the manifold and is bounded by Theorem 3.

K Factual and Counterfactual Loss Ablations

Tables 6 and 7 show the loss term ablation results for all factual loss terms in GLIDE. Tables 8 and 9 show the loss term ablation results for all counterfactual loss terms in GLIDE.

Each loss supports a distinct aspect of the method and removing any one lowers performance somewhere, confirming all are necessary rather than redundant. Removing a counterfactual loss consistently lowers CF Validity, and removing a factual loss consistently lowers Fid, so local manifold

Table 8: Counterfactual validity results across datasets and loss ablations. All loss terms are important. Full GLIDE loss has the best performance.

Dataset	Full GLIDE Loss	-$\mathcal{L}_{adj}$	-$\mathcal{L}_{entropy}$	-$\mathcal{L}_{margin}$	-$\mathcal{L}_{ortho}$	-$\mathcal{L}_{prox}$
AIDS	$\mathbf{0.97 \pm 0.05}$	0.49 ± 0.35	0.95 ± 0.05	0.54 ± 0.32	0.89 ± 0.07	0.71 ± 0.31
BA_2Motifs	$\mathbf{1.00 \pm 0.00}$	0.68 ± 0.45	0.71 ± 0.41	1.00 ± 0.00	0.83 ± 0.24	0.50 ± 0.41
MUTAG	$\mathbf{0.99 \pm 0.00}$	0.77 ± 0.04	0.42 ± 0.01	0.63 ± 0.14	0.53 ± 0.16	0.66 ± 0.16
NCI1	$\mathbf{0.88 \pm 0.01}$	0.20 ± 0.01	0.21 ± 0.01	0.20 ± 0.01	0.20 ± 0.01	0.21 ± 0.01
PROTEINS	$\mathbf{0.91 \pm 0.14}$	0.81 ± 0.16	0.66 ± 0.03	0.84 ± 0.13	0.74 ± 0.12	0.75 ± 0.15

Table 9: Counterfactual proximity results across datasets and loss ablations. All loss terms are important. Full GLIDE loss has the best performance.

Dataset	Full GLIDE Loss	$-\mathcal{L}_{adj}$	$-\mathcal{L}_{entropy}$	$-\mathcal{L}_{margin}$	$-\mathcal{L}_{ortho}$	$-\mathcal{L}_{prox}$
AIDS	$\mathbf{3.11 \pm 0.31}$	3.27 ± 0.16	3.48 ± 0.22	3.28 ± 0.13	3.25 ± 0.22	3.43 ± 0.09
BA_2Motifs	$\mathbf{2.74 \pm 1.94}$	4.54 ± 0.24	4.63 ± 0.06	4.71 ± 0.33	4.46 ± 0.08	4.57 ± 0.28
MUTAG	$\mathbf{2.07 \pm 0.21}$	2.02 ± 0.09	2.00 ± 0.06	2.93 ± 0.07	2.00 ± 0.14	2.03 ± 0.02
NCI1	$\mathbf{1.38 \pm 0.12}$	1.76 ± 0.37	2.08 ± 0.11	1.54 ± 0.13	1.54 ± 0.09	2.18 ± 0.29
PROTEINS	$\mathbf{1.30 \pm 1.84}$	1.61 ± 0.07	1.59 ± 0.09	1.59 ± 0.02	1.55 ± 0.05	1.69 ± 0.05

consistency (tangency, closure, orthogonality, proximity) is doing real work, not just adding regularization noise. For counterfactuals, entropy is most important for exploration/escaping local minima (its removal produces some of the largest validity drops, e.g. MUTAG); orthogonality stabilizes optimization and enforces the independence guarantee of Proposition 2; proximity removal hurts both CF proximity and CF validity together, consistent with it having dual function as an optimization stabilizer, not just a post-hoc bound; adjoint-loss removal gives the most consistent large validity drops, direct empirical evidence that the reductive decomposition of Theorem 1 is practically load-bearing, not just theoretically elegant.

L Symmetry discovery via Lie algebra parameterization

As mentioned in the main paper, a key distinction of our work is supervised disentanglement. We highlight main differences in our approach: (i) Augerino Benton et al. (2020) and LieGAN Yang et al. (2023) optimize unsupervised objectives (invariance of the augmentation distribution or the data distribution, respectively), making them insensitive to the classifier's decision boundary; our margin-based orbit loss derives the learning signal directly from the prediction task. (ii) LaLiGAN Yang et al. (2024) operates in latent space as we do, but learns a single undifferentiated symmetry group for generative reconstruction rather than the factual/counterfactual decomposition that explainability requires. (iii) LieGG Moskalev et al. (2023) extracts symmetries post-hoc from a frozen model, whereas our generators are trained jointly with the encoder in an ante-hoc manner and are the primary object of interest rather than a diagnostic tool. We therefore occupy a distinct position. We are the first framework to learn a disentangled Lie algebra decomposition supervised by classification margin, and to deploy the two resulting subalgebras as factual and counterfactual explanations of GNN predictions.

L.1 LieGAN and Augerino+ as Factual Explainers

LieGAN Yang et al. (2023) and Augerino^{+} Benton et al. (2020) are methods for symmetry-discovery. We adapt them for explainability.

L.2 Shared adaptation: latent-space transfer.

Rather than operating on raw graphs, we apply both baselines in the latent space of f_θ. After the encoder warmup stage (Stage 0), we extract embedding–label pairs $\{(z_i, y_i)\}$ where $z_i = f_\theta(G_i) \in \mathbb{R}^d$, and train each baseline on this dataset. Both baselines learn a skew-symmetric generator basis $\{A_{f,k^*}\}_{k=1}^{K}$ and are augmented with the same ℓ_1 sparsity penalty $\lambda_{\text{sp}}\|A_{f,k^*}\|_1$ and Lie algebra regularization $\lambda_{\text{Lie}}\,\mathcal{R}_{\text{Lie}}$ used in Phase 1, enabling a controlled comparison of inductive biases alone.

L.2.1 LIEGAN.

LieGAN trains one generator per class $c \in \mathcal{Y}$ via an adversarial objective: the generator learns to transform embeddings such that a discriminator cannot distinguish transformed from real class-c samples. The resulting generator parametrizes the orbit map

$$z \;\mapsto\; \exp\Big(\sum_k \theta_k \, A_{f,k^*}\Big) z, \qquad \theta_k \sim \mathcal{N}(0, \sigma^2), \tag{43}$$

which constitutes LieGAN's factual explanation for class c.

L.2.2 AUGERINO$^+$.

Augerino additionally learns a distribution over group coefficients $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$ alongside the generator basis, maximizing the invariance of the classifier's output under samples from that distribution. The training objective is

$$\mathcal{L}_{\mathrm{Aug}^+} \;=\; \mathbb{E}_{\boldsymbol{\theta}}\big[\mathcal{L}_{\mathrm{CE}}\big(f_c(z'),\, y\big)\big] \;+\; \lambda_{\mathrm{Lie}}\, \mathcal{R}_{\mathrm{Lie}} \;+\; \lambda_{\mathrm{sp}}\, \|A_{f,k^*}\|_1 \;-\; \tfrac{1}{2} \textstyle\sum_k \log \sigma_k, \tag{44}$$

where the final term is an entropy bonus that encourages the distribution to widen. At evaluation time the factual orbit is computed deterministically at $\boldsymbol{\theta} = \boldsymbol{\mu}$: $z' = \exp\big(\sum_k \mu_k A_{f,k^*}\big) z$.

L.2.3 UNIFIED EVALUATION.

The generators learned by each baseline are injected into GLIDE in place of the factual generator parameters and evaluated with four metrics. Let $\boldsymbol{t} = \mathrm{coef_net}([z,\, \mathbf{e}_{\bar{y}}])$ denote the predicted generator coefficients, and $k^* = \arg\max_k |t_k|$ the index of the top (most active) generator. We define two complementary fidelity scores following the sufficiency/necessity decomposition of Yuan et al. (2021). Fid^- (sufficiency) measures the mean target-class probability when only the top generator is applied,

$$\mathrm{Fid}^- \;=\; \mathbb{E}_G\Big[p_{\bar{y}}\big(f_c(z^-)\big)\Big], \quad z^- = g\big(\boldsymbol{t}^-\big) z, \quad t_k^- = \begin{cases} t_k & k = k^* \\ 0 & \text{otherwise,} \end{cases} \tag{45}$$

quantifying how much of the factual explanation is carried by the single most explanatory generator. Fid^+ (necessity) measures the drop in target-class probability when the top generator is removed,

$$\mathrm{Fid}^+ \;=\; \mathbb{E}_G\Big[p_{\bar{y}}\big(f_c(z)\big) - p_{\bar{y}}\big(f_c(z^+)\big)\Big], \quad z^+ = g\big(\boldsymbol{t}^+\big) z, \quad t_k^+ = \begin{cases} 0 & k = k^* \\ t_k & \text{otherwise,} \end{cases} \tag{46}$$

quantifying how much the explanation degrades without the top generator. Here $g(\boldsymbol{t}) = \exp\big(\sum_k t_k A_{f,k^*}\big)$ and $p_{\bar{y}}(f_c(\cdot))$ denotes the softmax probability of class $\bar{y}$. Accuracy reports the standard classification accuracy of f_c on the original embeddings, confirming encoder parity across methods. Proximity measures the mean ℓ_2 displacement $\mathbb{E}[\|z' - z\|_2]$, and subspace alignment

$$\mathrm{Align}(A, B) \;=\; \frac{1}{r} \sum_{i=1}^{r} \cos\theta_i(A, B), \qquad r = \min(\mathrm{rank}(A),\, \mathrm{rank}(B)), \tag{47}$$

measures whether the two methods recover the same symmetry directions.

L.2.4 KEY DISTINCTION FROM GLIDE.

Both baselines discover symmetries without an explicit coupling to the classifier's decision boundary. LieGAN maximizes adversarial realism and Augerino$^+$ minimizes the task loss on augmented samples, encouraging the classifier to be indifferent to the transformation. Our Phase 1 instead directly maximizes the classification margin at the transformed embedding,

$$\mathcal{L}_{\mathrm{orbit}} \;=\; \mathbb{E}\Big[\max\big(0,\; \gamma - \Delta_f(z', y)\big)\Big] \;+\; \lambda_{\mathrm{Lie}}\, \mathcal{R}_{\mathrm{Lie}}, \tag{48}$$

actively pulling transformed embeddings into the class region rather than flattening the loss surface around z or maximizing realism. An orbit that maximizes adversarial realism or invariance need not stay within the class region; one trained to preserve margin is explicitly penalized for leaving it. This distinction is directly measurable via Fid^+, and results for all methods are reported in Table 1.

Table 10: Degeneracy analysis for unsupervised factual explanation baselines on BA-2Motifs. Both LieGAN and Augerino$^+$ learn near-zero generator matrices, producing group elements within 10^{-3} of the identity and orbit radii below 10^{-3}. The resulting $\mathrm{Fid}^- \approx 1.0$ reflects the trivial fact that the identity transformation preserves every class label, not a meaningful factual explanation. GLIDE avoids this collapse by construction as the margin loss requires the orbit to move z into (and explore) the class manifold.

Metric	**LieGAN**	**Augerino$^+$**	**GLIDE**
Generator geometry			
Mean $\|A_{f,k^*}\|_F$ (↑)	0.4271 ± 0.1348	0.4530 ± 0.0825	$\mathbf{0.9837 \pm 0.0176}$
Top-generator weight fraction (↑)	0.0402 ± 0.0008	0.0219 ± 0.0001	$\mathbf{0.2561 \pm 0.0843}$
Orbit triviality			
Mean $\|g(\boldsymbol{t})z - z\|_2$ (↑)	0.0026 ± 0.0010	0.0000 ± 0.0000	$\mathbf{4.4411 \pm 0.4365}$
Mean $\|\exp(\sum_k t_k A_{f,k^*}) - I\|_F$ (↑)	0.0049 ± 0.0018	0.0000 ± 0.0000	$\mathbf{1.4862 \pm 0.0779}$
Factual explanation quality			
$\mathrm{Fid}^- \uparrow$	0.9984 ± 0.0010	0.9957 ± 0.0017	$\mathbf{0.8546 \pm 0.0521}$
$\mathrm{Fid}^+ \downarrow$	0.0000 ± 0.0000	0.0000 ± 0.0000	$\mathbf{0.0000 \pm 0.0000}$

Note: $\|A_{f,k^*}\|_F$ is the mean Frobenius norm of learned generator matrices. $\|g(\boldsymbol{t})z - z\|_2$ is the mean ℓ_2 orbit radius in latent space. $\|\exp(\cdot) - I\|_F$ measures distance of the group element from the identity. Top-generator weight fraction is $\max_k |t_k| / \sum_k |t_k|$; a value near 1 indicates a single generator dominates.

L.3 Degeneracy of LieGAN and Augerino+ Baselines

LieGAN Yang et al. (2023) and Augerino$^+$ Benton et al. (2020) are symmetry-discovery methods, not explainers. We adapted them for comparison since no other symmetry-based explainers exist; however, as Table 3 shows, their fidelity metrics reflect degenerate solutions rather than meaningful explanations, since neither method is trained with a downstream classification objective. Table 10 shows that even though the fidelity metrics for these methods are desirable at surface level, they are actually degenerate/ trivial solutions. This naturally arises because these methods are not trained with the downstream classification objective; they are frameworks for symmetry-discovery as previously stated.

M Counterfactual Baseline Details

We describe specifically how we implemented and/or adapted each counterfactual explainer baseline in the following section.

M.1 Shared metrics.

CF Fidelity measures the fraction of test graphs for which a valid counterfactual is produced,

$$\text{CF-Fid} \;=\; \mathbb{E}_G\big[\mathbf{1}\big[f_\theta(G') \neq f_\theta(G)\big]\big]. \tag{49}$$

CF Proximity measures the mean normalized graph edit distance,

$$\text{CF-Prox} \;=\; \mathbb{E}_G\left[\frac{\|A_G - A_{G'}\|_1}{2n^2}\right], \tag{50}$$

where $n = |V(G)|$; lower values indicate smaller perturbations. CF Sparsity measures the fraction of edges that are not perturbed,

$$\text{CF-Spar} \;=\; \mathbb{E}_G\left[1 - \frac{\|A_G - A_{G'}\|_1/2}{|E(G)|}\right], \tag{51}$$

with higher values indicating more targeted edits. For GLIDE, G' is defined implicitly via the latent transformation $z' = \exp\big(\sum_k t_k A_{f,k^*}\big) z$, so CF Proximity and Sparsity are computed in embedding space as $\mathbb{E}[\|z' - z\|_2]$ and reported alongside input-space baselines for comparison.

M.2 SHARED ADAPTATION.

In each case we make the minimal adaptation required. The internal GNN is replaced by a frozen wrapper around f_θ that exposes whatever interface the baseline requires, so that the same model is being explained across all methods.

M.3 CLEAR.

CLEAR Ma et al. (2022) generates counterfactual graphs via a VAE that conditions on a causal confounder u. Our datasets have no native confounder, so we run the u-free ablation (equivalent to CLEAR's CVAE variant), setting $u = \mathbf{0}$. The training objective is a three-term ELBO,

$$\mathcal{L}_{\text{CLEAR}} \;=\; D_{\text{KL}}\big(q(z \mid G, \bar{y}) \,\|\, p(z)\big) + \beta_x \, d_x(X, \hat{X}) + \beta_a \, d_a(A, \hat{A}) + \alpha \, \mathcal{L}_{\text{CE}}\big(f_\theta(G'),\, \bar{y}\big), \quad (52)$$

where the CF loss term is activated after an initial warm-up phase to allow the VAE to stabilize before the prediction signal is introduced.

M.4 GCFEXPLAINER.

GCFExplainer Kosan et al. (2022) produces a global counterfactual summary $\mathcal{S} = \{G'_1, \ldots, G'_k\}$ via an importance-guided vertex reinforced random walk over the graph edit space, rather than one CF per input. We evaluate it in two modes. For per-instance metrics, each test graph G_i is assigned its nearest summary graph by embedding distance,

$$G'_i \;=\; \arg\min_{G' \in \mathcal{S}} \big\| f_\theta(G_i) - f_\theta(G') \big\|_2, \quad (53)$$

after which CF Fidelity, Proximity, and Sparsity are computed as above. We additionally report summary-level coverage,

$$\text{Coverage} \;=\; \frac{1}{|\mathcal{G}_0|} \sum_{G \in \mathcal{G}_0} \mathbf{1}\big[\min_{G' \in \mathcal{S}} d(G, G') \le \theta\big], \quad (54)$$

measuring the fraction of class-0 input graphs covered by the summary within distance threshold θ, and mean frequency, which measures how consistently the random walk reinforces the same counterfactuals.

M.5 D4EXPLAINER.

D4Explainer Chen et al. (2023) uses a score-based diffusion model to denoise perturbed adjacency matrices toward counterfactual graphs. The training loss combines a weighted BCE denoising term and a counterfactual prediction term,

$$\mathcal{L}_{\text{D4}} \;=\; \mathcal{L}_{\text{diff}}(\sigma) \;+\; \alpha_{\text{cf}} \, \mathcal{L}_{\text{cf}}, \quad (55)$$

where σ is the noise level sampled during training. CF Sparsity is derived from D4Explainer's internal modification ratio $r_{\text{mod}} = \|A_G - A_{G'}\|_1 / 2|E(G)|$ as CF-Spar $= 1 - r_{\text{mod}}$, consistent with Eq. 51.

M.6 KEY DISTINCTION FROM GLIDE.

All four baselines generate counterfactuals in input space, producing a modified graph G' for each test instance (or a global summary in the case of GCFExplainer). This provides directly interpretable edge-level explanations but requires either graph reconstruction from a VAE (CLEAR Ma et al. (2022)), a global random walk (GCFExplainer Kosan et al. (2022)), or a separate diffusion network (D4Explainer Chen et al. (2023)). GLIDE instead applies a single learned group transformation in latent space, $z' = \exp\big(\sum_k t_k A_{f,k^*}\big) z$, which is shared across all test graphs of a given class and requires no graph reconstruction. The generator basis $\{A_{f,k^*}\}$ provides a class-level geometric explanation of which directions in latent space correspond to class membership, rather than a per-instance list of edge modifications. Results for all baselines are reported in Table 2.

Table 11: Counterfactual validity across datasets. Values are mean ± standard deviation over 3 runs. Higher is better.

Dataset	Ablation (1)	Ablation (2)	GLIDE
BA_2Motifs	0.21 ± 0.07	0.91 ± 0.12	$\mathbf{1.00 \pm 0.00}$
MUTAG	0.56 ± 0.15	0.68 ± 0.47	$\mathbf{0.99 \pm 0.00}$
PROTEINS	0.74 ± 0.01	0.82 ± 0.12	$\mathbf{0.91 \pm 0.14}$
NCI1	0.20 ± 0.06	0.19 ± 0.03	$\mathbf{0.88 \pm 0.01}$
AIDS	0.88 ± 0.03	0.89 ± 0.01	$\mathbf{0.97 \pm 0.05}$

Table 12: Counterfactual proximity across datasets. Values are mean ± standard deviation over 3 runs. Lower is better. * denotes that the CF proximity is trivially low, and as reflected in the CF validity score, does not produce valid counterfactuals.

Dataset	Ablation (1)	Ablation (2)	GLIDE
BA_2Motifs	4.68 ± 0.23	4.70 ± 0.04	$\mathbf{2.74 \pm 1.94}$
MUTAG	8.03 ± 0.14	8.50 ± 0.05	$\mathbf{2.07 \pm 0.21}$
PROTEINS	5.33 ± 0.06	$0.57 \pm 0.01^\star$	$\mathbf{1.30 \pm 1.84}$
NCI1	0.22 ± 0.01	$0.05 \pm 0.04^\star$	$\mathbf{1.38 \pm 0.12}$
AIDS	5.32 ± 0.08	5.73 ± 0.04	$\mathbf{3.11 \pm 0.31}$

M.7 Additional Ablations

We add two additional baselines: (1) A raw gradient-ascent counterfactual: $z = z + \nu \cdot \nabla_z(logit_{y_cf} logit_y)$, iterated with the same step budget as the coefficient network in GLIDE, with no group/ exponential-map structure and no generator basis; and (2) an ungrounded low-rank perturbation ablation of our own architecture: identical generator parameterization $A_k = U_k V_k^T$ and coefficient network g_ψ, but with the exponential-map update $z = exp(_k A_k)z$ replaced by a simple additive update $z = z +_k (A_k z)$, and with L_{ortho}/L_{adj} (the closure/ reductive-structure losses) removed. Both run with encoder and factual generators frozen, all five datasets. This will isolate the contribution of the group-exponential parameterization itself and the reductive-decomposition/ closure losses, while holding the low-rank architecture fixed Both perform substantially worse than GLIDE (Tables 11 and 12), providing direct empirical support for the necessity of the Lie-theoretic structure.

The results are presented in Tables 11 and 12. (1) shows significantly worse proximity (2-10x) and (2) a degraded validity/proximity trade-off vs. full GLIDE, giving empirical support for the necessity of the Lie-theoretic structure over any low-rank latent method.

N Empirical Validation Details

This appendix provides the complete experimental protocols, extended analyses, and diagnostic results supporting the compressed validation summary in Section 7.

N.1 Validation of Proposition 1 Assumptions

We empirically validate all three assumptions for Proposition 1. Below we provide testable diagnostics for each of the conditions, empirically validating that they are satisfied in our setting.

Non-degeneracy ($\nabla_z y(z) \neq 0$): Across all datasets, there are no degenerate test samples and across datasets, the mean gradient norm is listed in Table 13, confirming that this condition always holds in the empirical setting.

Faithfulness (injectivity of $\rho_z : A \to Az$): The mean empirical rank of the Jacobian $[A_1 z, \cdots, A_K z]$ across all test samples for all datasets is full-rank (k=4; Table 13).

Table 13: Assumption checks across three independent runs, empirically validating Proposition 1 assumptions on all datasets. Non-degeneracy reports the mean gradient norm $\|\nabla f\|$, full rank reports the fraction of samples satisfying the full-rank condition, local transitivity reports the fraction satisfying the local transitivity condition, and mean $\alpha^\star$ reports the mean normalized boundary displacement.

Dataset	Non-degeneracy: $\mathrm{mean}(\|\nabla f\|)$	Local transitivity	Mean $\alpha^\star$
BA-2Motifs	1.7110 ± 0.0515	$100.0\% \pm 0.0\%$	0.198 ± 0.049
MUTAG	4.0412 ± 0.2642	$98.9\% \pm 0.7\%$	0.492 ± 0.019
PROTEINS	2.6152 ± 0.0842	$75.6\% \pm 3.8\%$	0.186 ± 0.120
NCI1	3.4943 ± 0.0267	$88.9\% \pm 1.7\%$	0.321 ± 0.454
AIDS	2.6152 ± 0.0842	$94.6\% \pm 2.1\%$	0.463 ± 0.015

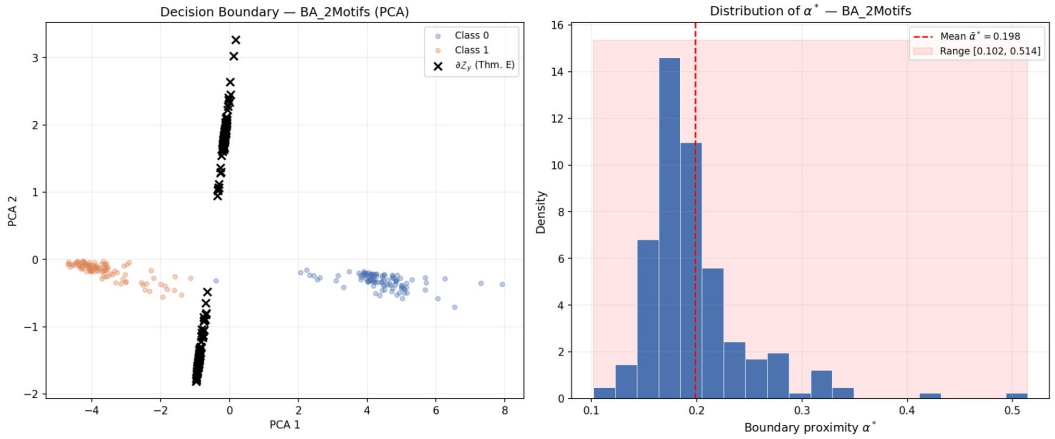

Figure 7: **Decision boundary as counterfactual leaf space.** *Left:* PCA projection with recovered boundary points $\partial\mathcal{Z}_y$ (black crosses, $n = 200$) confirms Thm. 2. *Right:* Distribution of α^* (mean 0.198, range $[0.102, 0.514]$); the smooth, unimodal shape confirms leaves vary continuously rather than clustering discretely, consistent with the smooth foliation of Theorem 2.

Local transitivity ($\rho_z(\mathfrak{h}_z = T_z Z_y)$: In the Appendix N.2 surface-check protocol tests, for every recovered boundary point $z^\star$, whether $\hat{y}(z^\star + \epsilon v) \neq \hat{y}(z^\star - \epsilon v)$ along the dominant generator direction. This is a necessary condition for $z^\star$ to lie on δZ_y rather than merely near it. All 200 recovered points on BA-2Motifs pass this test (100%). We repeated this test across all datasets and obtained the following results: the transitivity pass rate for MUTAG is 98%, for AIDS 94%, for NCI1 88%, and for PROTEINS 73%. The PROTEINS pass rate (73%) is lower, reflecting genuinely irregular decision-boundary geometry in that dataset. This is consistent with PROTEINS being the hardest classification task among the datasets. * remains low across all datasets. Therefore, the underlying local transitivity assumption holds empirically (see Appendix N.2).

N.2 Validation of Theorem 2: Full Boundary Analysis

Binary search procedure. For each test sample z we apply the trained coefficient network $g_\psi(z, y_{\text{cf}})$ to obtain the full predicted transformation matrix $B = \sum_k t_{cf,k} A_{cf,k}$. We then binary search over $\alpha \in [0, 1]$ (20 iterations, tolerance 10^{-4}) to find the minimal α^* such that $\hat{y}(\exp(\alpha^* B) \cdot z) = y_{\text{cf}} \neq \hat{y}(z)$. The point $z^* = \exp(\alpha^* B) \cdot z$ is treated as a sample on the decision boundary $\partial\mathcal{Z}_y$. We apply this procedure to all 200 BA-2Motifs test samples and project the resulting boundary points into two dimensions via PCA alongside the original embeddings.

Surface-check protocol. To verify that recovered points lie on the boundary surface rather than merely near it, we apply the following test to each z^*: let $v = A_{cf,k^*} z^* / \|A_{cf,k^*} z^*\|_2$ be the dominant generator direction, and check whether $\hat{y}(z^* + \epsilon v) \neq \hat{y}(z^* - \epsilon v)$ for $\epsilon = 0.05$. A point on $\partial\mathcal{Z}_y$

must satisfy this condition; a point that drifted close to the boundary by chance need not. All 200 recovered points pass this test (100% verification rate), providing direct empirical evidence that the counterfactual leaves trace $\partial\mathcal{Z}_y$ exactly as Theorem 2 claims.

Extended analysis of α^* distribution. The distribution of α^* values (Figure 7, right) is strongly right-skewed and tightly concentrated: the mode falls near $\alpha^* \approx 0.18$, over 80% of samples satisfy $\alpha^* \leq 0.30$, and the hard ceiling at $\alpha^* = 0.514$ confirms no sample requires more than half the full predicted transformation to reach the boundary. This structure admits a clean geometric interpretation under Theorem 2: leaves near the boundary correspond to small α^* values, while leaves deeper in the class interior correspond to larger α^*. The smooth, unimodal shape confirms these leaves are continuously nested rather than discretely clustered, consistent with the smooth foliation guaranteed by the Frobenius theorem (Theorem 2 proof). If counterfactual flows were geometrically arbitrary rather than constrained by Lie group structure, the boundary samples would be distributed diffusely across the latent space; the observed geometric coherence is a direct consequence of all trajectories being confined to the same algebraic manifold.

Results on all datasets. We report in Table 13 the mean $\alpha^\star$ (Theorem 2 validation) for all datasets, showing that the underlying local transitivity assumption holds empirically. Please see Appendix N.1 for full discussion of $\alpha^\star$.

N.3 Validation of Theorem 3: Manifold Proximity Guarantees

Theorem 3 establishes a static upper bound on the latent displacement $\|z' - z\|_2$ as a function of generator norms and coefficient magnitudes. Theorem 5 strengthens this to a dynamic guarantee proving that the bound tightens monotonically over training. We provide the full quantitative analysis of the proximity bound here.

The boundary analysis provides a direct empirical window onto these guarantees. The scaling factor α^* is a normalized measure of the transformation magnitude required to reach the boundary: $\alpha^* = 1$ corresponds to the full predicted transformation B, whose proximity to z is measured by the effective B-norm. On BA-2Motifs, the effective B-norm is 8.845 and the mean z-norm is 3.282. Applying Theorem 3 with these values gives an upper bound on $\|z^* - z\|_2$ for the boundary point. Since $\bar{\alpha}^* = 0.198$, the observed displacement at the boundary is approximately 19.8% of the full transformation, well within the theoretical envelope. For MUTAG, the mean boundary displacement is $\approx 49.2\%$ of the full transformation; PROTEINS is $\approx 18.6\%$; NCI1 is $\approx 32.1\%$; and AIDS is $\approx 46.3\%$.

For all test samples, $\alpha^* \leq 0.514$, meaning no sample requires more than 51.4% of the full transformation to reach the boundary. This is strong empirical evidence that the counterfactual flows remain geometrically local and they do not require large movement in latent space to achieve a label flip, which is the key failure mode of unconstrained counterfactual methods.

We directly plot the proximity over epochs. Figure 8 shows the expected decline, again validating Theorem 3 and Theorem 5.

Remark 1. *The limiting bound $\|z\|_2(e^{M^*\|t\|_1} - 1)$ approaches zero if and only if $M^*\|t\|_1 \to 0$. This does not require generator collapse ($M^* = 0$, which $\mathcal{L}_{\text{margin}}$ prevents); it can occur if the coefficient network g_ψ learns to predict vanishing $\|t\|_1$ as generators align more precisely with the decision boundary normal, requiring smaller steps to achieve the label flip. Empirically (Figure 8), OOD drift decays to near-zero while CF Fidelity remains 1.000, confirming that $M^* > 0$ and $\|t\|_1 \to 0$ simultaneously as the generators become geometrically efficient rather than large.*

N.4 Theorem 6: Spectral Validation Details

Experimental setup. To isolate generator convergence dynamics from encoder dynamics, the encoder f_θ is trained to convergence in Stage 0 and then frozen before Phase 1 begins; generators are reinitialized to random matrices prior to starting Phase 1. Freezing the encoder is essential because it ensures that the orbit loss curvature reflects only the generator dynamics, so fitted exponential decay rates are not contaminated by the simultaneous convergence of the encoder.

Staged training dynamics. Figure 9 shows the complete training timeline. During Stage 0 (grey), encoder test accuracy rises monotonically from 0.50 to 1.00; generators are inactive throughout

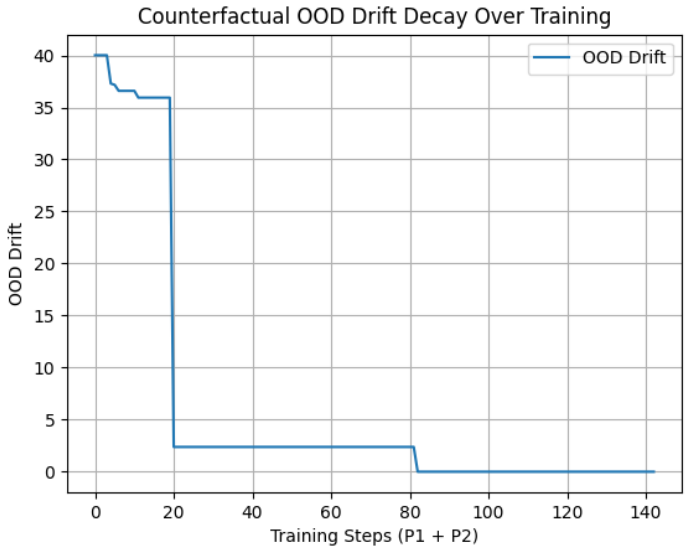

Figure 8: Example plot on BA-2Motifs. OOD drift (proximity) decreases over training, empirically validating Theorem 3 and Theorem 5.

(orbit radius identically zero, bottom panel). Fid^- quickly increases because the encoder already separates the two classes almost perfectly, so the factual generators satisfy the margin constraint within the first epoch. In Phase 2, CF Fidelity rises from to 1.000. The CF generators converge rapidly because the latent space is, by this point, highly organized. The bottom panel further confirms Theorem 3, showing that the factual orbit radius contracts over Phase 1 as generators converge to the minimum-norm orbit satisfying the margin constraint, while the CF orbit radius grows during Phase 2 as counterfactual generators learn to traverse the decision boundary.

Joint-training ablation. Figure 10 validates the curvature ordering of Theorem 6 without any explicit phase separation. When both generator sets are trained simultaneously from the same frozen encoder, Fid^- converges at epoch 0 while CF Fidelity does not cross 0.85 until epoch 10. The ordering is not imposed by the training protocol but arises naturally from the spectral gap as the measured Hessian curvatures are $\lambda_f = 1086$ and $\lambda_{cf} = 475$, confirming the inequality $\lambda_f > \lambda_{cf}$ predicted by Theorem 6. We note that joint training is acceptable in this ablation precisely because $\Delta = \lambda_f - \lambda_{cf}$ is moderate; as discussed in Section 5.3, when Δ is large the cross-subspace interference term (Equation 29) becomes non-negligible, and explicit phase separation is necessary to avoid degraded CF convergence.

N.5 Theorem 4: Generator Diagnostic Details

Generator norm concentration. After training with the winner-take-all entropy loss and hard Gumbel-Softmax masking (applied after epoch 50), for BA-2Motifs, generator A_0 achieves Frobenius norm $\|A_0\|_F = 0.2447$ while all remaining generators have norm ≈ 0. The coefficient network correspondingly concentrates all magnitude on A_0 (mean coefficient 13.56, maximum 16.04 across test samples), with the hard mask enforcing the one-hot selection that Theorem 4 identifies as the minimum-ℓ_2-norm solution.

For MUTAG, generator A_0 achieves Frobenius norm $\|A_0\|_F = 0.1046$ while all remaining generators have norm ≈ 0. The coefficient network correspondingly concentrates all magnitude on A_0 (mean coefficient 12.54, maximum 14.12 across test samples).

For NCI1, generator A_0 achieves Frobenius norm $\|A_0\|_F = 0.0348$ while all remaining generators have norm ≈ 0. The coefficient network correspondingly concentrates all magnitude on A_0 (mean coefficient 135.6, maximum 160.4 across test samples).

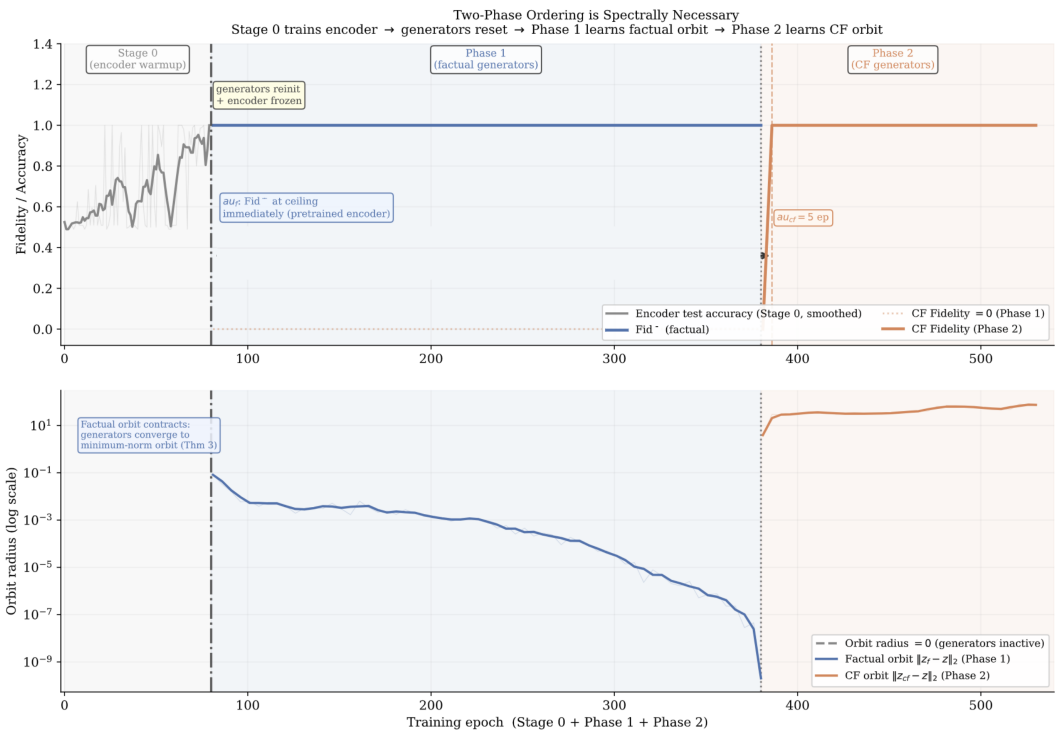

Figure 9: Empirical validation of Theorem 6 (Two-Phase Ordering is Spectrally Necessary). *Top:* Stage 0 (grey) trains the encoder to perfect accuracy; generators are the reinitialized and the encoder frozen (dash-dot line). Phase 1 (blue) trains factual generators: Fid^+ reaches its ceiling immediately ($\tau_f \approx 0$, given the pretrained encoder). Phase 2 (orange) trains CF generators: CF Fidelity rises to 1.000 within $\tau_{cf} = 5$ epochs. *Bottom:* Factual orbit radius contracts during Phase 1 as generators converge to the minimum-norm orbit satisfying the margin (Theorem 3); CF orbit radius grows in Phase 2 as counterfactual generators learn to traverse the decision boundary.

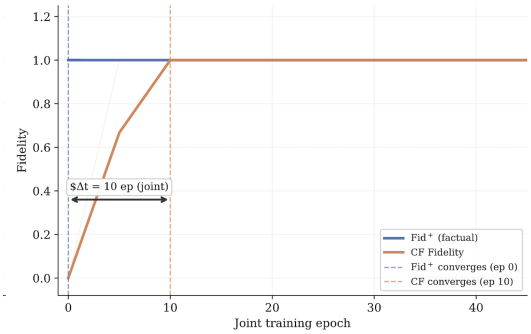

Figure 10: Further empirical validation of Theorem 6. Both generators trained simultaneously on the same frozen encoder. Fid^+ converges at epoch 0; CF Fidelity converges at epoch 10. The staircase arises without explicit phase separation, driven purely by the spectral gap $\lambda_f > \lambda_{cf}$ identified in Theorem 6.

For PROTEINS, generator A_0 achieves Frobenius norm $\|A_0\|_F = 0.1657$ while all remaining generators have norm ≈ 0. The coefficient network correspondingly concentrates all magnitude on A_0 (mean coefficient 2.32, maximum 7.62 across test samples).

For AIDS, generator A_0 achieves Frobenius norm $\|A_0\|_F = 0.1074$. while all remaining generators have norm ≈ 0. The coefficient network correspondingly concentrates all magnitude on A_0 (mean coefficient 9.60, maximum 29.81 across test samples).

Top-1 sparsity test. As an independent verification, we evaluate CF Fidelity using only the top-ranked generator (zeroing all others) across all test samples. CF Fidelity $= 1.000$ is achieved with this single generator alone, directly confirming that one Lie algebra direction is both necessary and sufficient to characterize the full decision boundary crossing. Table 5 reports top-1 sparsity scores across all benchmark datasets. Generator selection statistics confirm that the winner-take-all mechanism converges to a dominant generator ($\geq 75\%$ selection rate for k^* across datasets), consistent with Theorem 4.

O LINEARIZED GRADIENT FLOW

Linearized Gradient Flow Consider gradient flow near a critical point θ^*:

$$\dot{\theta} = -\nabla \mathcal{L}(\theta).$$

Let $\delta\theta = \theta - \theta^*$. Linearizing around θ^* yields:

$$\dot{\delta\theta} = -H\delta\theta, \quad H = \nabla^2 \mathcal{L}(\theta^*).$$

Thus local training dynamics are governed by a linear ODE.

Block Decomposition via Symmetry From symmetry invariance, the Hessian commutes with the group action:

$$H\rho(g) = \rho(g)H.$$

By Schur's lemma, H decomposes over isotypic components:

$$\Theta = \bigoplus_{\lambda \in \widehat{G}} V_\lambda, \quad H = \bigoplus_{\lambda} H_\lambda,$$

where H_λ is the restriction of H to V_λ.

Each symmetry sector evolves independently:

$$\dot{\delta\theta}_\lambda = -H_\lambda \delta\theta_\lambda.$$

Spectral Solution Let the eigenvalues of H_λ be:

$$0 < \lambda_1^\lambda \leq \lambda_2^\lambda \leq \cdots.$$

Diagonalizing:

$$H_\lambda = Q_\lambda \Lambda_\lambda Q_\lambda^\top,$$

the solution is:

$$\delta\theta_\lambda(t) = e^{-H_\lambda t}\delta\theta_\lambda(0).$$

In eigenbasis coordinates:

$$\delta\theta_{\lambda,i}(t) = e^{-\lambda_i^\lambda t}\delta\theta_{\lambda,i}(0).$$

Thus each eigenmode decays exponentially.

Dominant Decay Rate The slowest decaying mode in V_λ corresponds to the smallest eigenvalue:

$$\lambda_{\min}(H_\lambda).$$

Hence asymptotically:

$$\|\delta\theta_\lambda(t)\| \sim e^{-\lambda_{\min}(H_\lambda)t}.$$

This defines the characteristic timescale:

$$\tau_\lambda = \frac{1}{\lambda_{\min}(H_\lambda)}.$$

Curvature Interpretation The Hessian measures local curvature of the loss surface:

$$\mathcal{L}(\theta^* + v) \approx \mathcal{L}(\theta^*) + \frac{1}{2} v^\top H v.$$

Large eigenvalues correspond to steep curvature directions. Small eigenvalues correspond to flat directions.

Steeper curvature induces faster exponential contraction under gradient flow.

Hierarchical Symmetry Learning Suppose two symmetry sectors exist:

$$V_{\text{local}}, \quad V_{\text{global}}.$$

If their minimal eigenvalues satisfy:

$$\lambda_{\text{local}} \gg \lambda_{\text{global}},$$

then:

$$\tau_{\text{local}} = \frac{1}{\lambda_{\text{local}}} \ll \frac{1}{\lambda_{\text{global}}} = \tau_{\text{global}}.$$

This means that local symmetry-breaking modes decay rapidly, while global symmetry-breaking modes decay slowly.

The system first contracts toward the subspace $\Theta^{G_{\text{local}}}$, and only later toward $\Theta^{G_{\text{global}}}$.

Spectral Gap Interpretation Define the spectral gap between sectors:

$$\Delta = \lambda_{\text{local}} - \lambda_{\text{global}}.$$

If Δ is large, contraction occurs in separated stages.

This multi-timescale decay produces the observed "staircase" symmetry learning:

$$\Theta \longrightarrow \Theta^{G_{\text{local}}} \longrightarrow \Theta^{G_{\text{global}}}.$$

Spectral separation of Hessian eigenvalues therefore directly explains hierarchical symmetry emergence.

P EXTENDING GLIDE TO MULTI-CLASS SETTING

All five benchmarks we provide are for binary classification, matching the baseline literature. The one-dimensional-complement assumption (dim $\mathfrak{m}_z = 1$, Thm. 4 proof, Appendix F.6, Step 1) is specific to binary classification, where the decision boundary is a single co-dimension-1 hypersurface with one transverse direction. For K-classes, a boundary point generically has up to K1 independent transverse directions, therefore, $\mathfrak{m}_z$ would generically become (K1)-dimensional, and Theorem 4's single-generator conclusion needs a per-target-class rule instead. However, not all of our results depend on the binary setting. Theorem 1's reductive decomposition and Theorem 6's spectral curvature argument do not use dim $\mathfrak{m}_z = 1$ and extend to the multi-class case unchanged. Proposition 1 and Theorem 4 specifically require revisiting. This is a promising direction for extending our work.

Q GUARANTEES AND BOUNDS ON CONTINUITY AND DIFFERENTIABILITY

On real-analytic (C^ω) smoothness. The Lie-algebraic action via the matrix exponential $\exp(X) = \sum_{k=0}^{\infty} \frac{X^k}{k!}$ does not merely provide C^1 or C^k differentiability, rather it is real-analytic (C^ω) over the matrix Lie algebra $\mathfrak{g} \subset \mathbb{R}^{d\times d}$. Because $g(t) = \exp(tX)$ is C^ω, the mapping $(X, z) \mapsto \exp(X)z$ is infinitely differentiable (C^∞) with respect to both the generator coefficients in X and the latent embedding z. Unlike L_0 graph masking, prototype sampling, or Gumbel-Softmax with Straight-Through Estimators (STEs) which introduce biased or non-existent higher-order gradients, the matrix exponential provides smooth, unbiased gradient flow across the entire latent space $\mathcal{Z}$.

Geometric Level-Set Guarantees (Sard's Theorem Transversality). Flow maps are C^∞. Therefore, we have:

- Smooth orbit submanifolds: The factual orbit $\mathcal{O}_{\hat{y}(z)} = \{\exp(X_{\text{fact}})z \mid X_{\text{fact}} \in \mathfrak{h}_z\}$ is a smooth C^∞ sub-manifold of $\mathcal{Z}$.
- Clean Boundary Transversality: By Sard's Theorem and Transversality theory, counterfactual flows generated by $X_{\text{cf}} \in \mathfrak{m}_z$ intersect the decision boundary $\partial\mathcal{Z}_y$ transversely. This rules out pathological or non-deterministic boundary crossings during gradient descent.

Quantitative Continuity Perturbation Bounds. To guarantee that continuous deformation does not push representations off the data manifold, the perturbation distance in latent space is strictly bounded by the Lie algebra operator norm: $\|\exp(X)z - z\|_2 \leq \|z\|_2 \left(e^{\|X\|_2} - 1\right)$.

For small generator norms $\|X\|_2 \leq M$, this yields a Lipschitz continuity guarantee: $\|\exp(X)z - z\|_2 \leq \|z\|_2 \cdot M \cdot e^M$.

This bound proves that the continuous mapping cannot produce discontinuous jump artifacts in latent space.

A note on Auto-Diffing. We would like to remark that auto-diffing does not invalidate or degrade the theoretical guarantees. Modern deep learning frameworks (e.g., PyTorch's torch.matrix_exp) compute $\exp(X)$ using the Higham Scaling-and-Squaring Algorithm with Padé Approximation: $\exp(X) \approx \left[P_p\left(2^{-s}X\right) Q_p\left(2^{-s}X\right)^{-1}\right]^{2^s}$, where P_p and Q_p are matrix polynomials of order p (typically $p = 13$).

From a theoretical and numerical perspective, empirical automatic differentiation through the matrix exponential fully preserves theoretical reliability across differentiability, gradient evaluation, and numerical stability. Theoretically, the matrix exponential $\exp(X) = \sum_{k=0}^{\infty} \frac{X^k}{k!}$ is real-analytic (C^∞). Empirical implementations evaluate this via a rational Padé approximation $P_p(X)Q_p(X)^{-1}$, which guarantees identical C^∞ smoothness across the entire domain provided Q_p remains invertible. Regarding gradient computation, algorithmic differentiation through the scaling-and-squaring graph evaluates the true analytical Fréchet derivative $d\exp_X(V) = \int_0^1 e^{(1-t)X} V e^{tX}\, dt$ exactly up to machine precision ($\approx 10^{-15}$). Finally, while theoretical conditioning depends on the operator norm

$\|X\|_2$, the empirical implementation maintains stability by dynamically scaling the generator via a factor 2^{-s} such that $\|2^{-s}X\|_2 < 1$, effectively preventing gradient explosion or vanishing during repeated matrix powering.

R COMPUTER RESOURCES

All experiments were implemented using `PyTorch Geometric` and run on either CPU or NVIDIA H200 GPUs. All datasets run through the full GLIDE pipeline (Phase 0 $\rightarrow$ Phase 1 Factual $\rightarrow$ Phase 2 Counterfactual) with a single run taking approximately 10 minutes. We run over 10 random seeds, compiling to a total of $8 - 9$ hours.